Health-state valuation using discrete choice models

Anna Nicolet Selivanova
2018

Health-state valuation using discrete choice models

Anna Nicolet Selivanova PhD Thesis
University of Groningen
University Medical Center Groningen

ISBN: 978-94-034-0862-0 (Printed version)
ISBN: 978-94-034-0861-3 (Electronic version)

Layout and design by: Anouk Westerdijk, persoonlijkproefschrift.nl
Printed by: Ipskamp Printing, proefschrift.net

university of groningen

Research Institute SHARE

university of groningen

Health-state valuation using discrete choice models

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 24 September 2018 at 11.00 hours

by

Anna Nicolet Selivanova

born on 27 March 1992
in Rjazan, Russia

Supervisors
Prof. E. Buskens
Dr. P. F. M. Krabbe

Assessment committee
Prof. M. J. Postma
Prof. C. D. Dirksen
Prof. J. J. V. Busschbach

TABLE OF CONTENTS

GENERAL INTRODUCTION

GENERAL INTRODUCTION

There are various definitions of health. Brüssow [1] made several attempts to define health but found it difficult, since the field of medicine is more interested in disease than health. The Oxford Living Dictionary of World English recently defined health as 'the state of being free from illness and injury'. But back in 1948 the World Health Organization defined it as 'a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity'. The comprehensive scope of the WHO definition is under debate, however [1, 2]. Nowadays, taking the WHO definition as their point of departure, many researchers focus on health-related quality of life (HRQoL), which has come to be regarded as an important outcome measure.

To be meaningful, a measure of HRQoL should assess not only the severity of a person's complaints or their occurrence, but also their impact. In other words, a measure of HRQoL should reflect how patients perceive or experience their own health status [3-7]. Various approaches are used to measure HRQoL. The one at the core of this thesis is the preference-based framework, which captures a person's overall health condition or health status in a single figure. Within that framework, several instruments (EQ-5D, HUI-3, SF-6D, AQol) have been developed, whereby 'preference' denotes the relative 'desirability' of a specific object. The measures obtained with preference-based methods are referred to as values. Those values can be used in health-outcomes research, disease-modeling studies, and economic evaluations for the comparison of different healthcare interventions and for the planning and monitoring of health programs. Within a preference-based measurement framework different health aspects (also called attributes) are weighted on the basis of assessments made by the respondent. These assessments are typically based on a comparison between health-state descriptions [8]. All preference-based methods require a comparative element in the judgmental task to elucidate the relative importance of the attributes. Another feature of preference-based measurement is that the respondents do not score the attributes one by one but consider the whole set of health attributes in their assessment [9].

An important question in health-state measurement is "Who should value health?", which raises an issue that has long been subject to heated debate. For the majority of instruments, the values for health states that are being used in health evaluations are derived from a representative community sample [10]. These generally healthy people are asked to judge hypothetical health states that are described by health attributes with certain levels of severity. Being tax payers, the general public are assumed not to serve own self-interest and, therefore, to embody principles of justice and equity. However, it is reasonable to assume that in many situations healthy subjects may be inadequately informed or lack sufficient imagination to make an appropriate judgment about the impact of hypothetical health states on their quality of life. Many researchers claim that individuals are the best judges of their own HRQoL. They are likely to be more adequately

informed than healthy people or more adept at imagining certain health states. Therefore, in the opinion of those researchers, it is the patients' judgments that should be elicited to obtain values for health states. That reasoning may be more compelling when the respondents have to take into account severely impaired health states, since people who have direct experience with impaired health may provide more reliable and valid health-state valuations [14].

Preference-based measures quantify multiple health attributes by condensing them into a single metric as a result of applying specific valuation techniques. The techniques commonly used for health-state measurement stem from the discipline of economics and are known to be complex and prone to biases [15, 16]. In fact, these techniques are becoming even more complicated through attempts to 'locate' death, i.e., to allow valuation by comparison to non-dead states and/or health states worse than death. These attempts to push beyond the quantification of health have sparked interest in methods that use cognitively less demanding tasks and that are firmly grounded in measurement theory. The most promising method in that regard is discrete choice modeling [17, 18].

The impetus for theoretical advances in discrete choice modeling has come largely from transportation planning, but the main body of research using choice modeling has been in the fields of marketing and economics. This technique requires participants to make choices among two or more scenarios (choice tasks) described by means of specific attributes with certain levels. Lately, interest in the use of choice models has increased in the field of health evaluation as well. Such models can further our understanding of how changes in specific health attributes influence preferences regarding a particular health state. All discrete choice models establish the relative merit of one phenomenon based on its relative attractiveness. Choice tasks are generally simple to complete, and they are often conducted without an interviewer through the form of postal or on-line surveys [19, 20].

The instruments that have been used for health-state measurements are known to have certain shortcomings; for instance, health values elicited from the general public are derived by methods that are complex and prone to bias. To overcome the shortcomings, we set out to assess the added value of an alternative approach, namely discrete choice modeling. Therefore, the aim of this thesis was to investigate the specific problems associated with preference-based measures of health states and with the methodology used to derive health-state values. More specifically this thesis sheds light upon the application of discrete choice modeling for measuring health states, with a special focus on EQ-5D health-state values. The **first chapter** covers the changes in phrasing and differences in valuation techniques in the EQ-5D instrument as a result of the introduction of the 5-level version alongside the current 3-level version. Specifically, a head-to-head comparison of the EQ-5D-3L and EQ-5D-5L was designed to explore differences in the health-state values produced by these two instruments using the discrete choice model. **The second chapter** investigates whether the inclusion of interactions between

various EQ-5D-3L health attributes (i.e., limited mobility or pain/discomfort may affect the appraisal of usual activities) leads to different values for health states, and whether a model with interactions would have better fit than a main-effects model. The **third chapter** considers whether people with experience of disease tend to assign different values to health states or more/less importance to certain health attributes than currently healthy respondents would do. The **fourth chapter** presents a separate study using a discrete choice model to determine the importance of certain criteria for new medical treatments. We explore whether there are differences in preference for these criteria between the general population and patients. The **fifth chapter** presents a study focused on a basic assumption in the valuation of health states, namely that respondents pay attention to all information in the health-state description and do not disregard information elements. For this investigation we used the eye-tracking technique.

REFERENCES

1. H. Brüssow. What is health? Microb Biotechnol. 2013; 6(4): 341–348.
2. Levine, S. The meaning of health, illness, and quality of life. Geggenmoose-Holzman I, Brenner H, Flick U, editors. Quality of life and health: concepts, methods and applications. Berlin: Blackwell Wissenschaft 1995; 7–12.
3. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. JAMA. 1994; 272(8): 619–626.
4. Testa M. Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. Med. Care. 2000; 38(9): 166- 174.
5. Bonomi AE, Patrick DL, Bushnell DM, Martin M. Validation of the United States' version of the World Health Organization Quality of Life (WHOQOL) instrument. J Clin Epidemiol.2000; 53(1): 1-12.
6. Sullivan M. The new subjective medicine: taking the patient's point of view on health care and health. Soc Sci Med. 2003; 56(7):1595–1604.
7. Hamming JF, De Vries J. Measuring quality of life. Br J Surg. 2007; 94: 923–4.
8. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: Health utility index. Pharmacoecon. 1995; 7: 503–520.
9. Fischer GW. Utility models for multiple objective decisions: do they accurately represent human preferences? Decis Sci. 1979; 10: 451–479.
10. Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the economic evaluation of health care programmes. Fourth ed. Oxford University Press; 2015.
11. Gandjour A. Theoretical foundation of patient v. population preferences in calculating QALYs. Med Decis Making 2010; 30 (4): 57-63.
12. Rand-Hendriksen K, Augestad L, Kristiansen IS, et al. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. Qual Life Res 2012; 21:1005–1012.
13. Neumann PJ, Ganiats TG, Russell LB, et al. eds. Cost-Effectiveness in Health and Medicine. Oxford University Press; 2016.
14. Jonker MF, Attema AE, Donkers B, et al. Are health state valuations from the general public biased? A test of health state preference dependency using self-assessed health and an efficient discrete choice experiment. Health Econ 2016; 1-14.
15. Doctor JN, Bleichrodt H, Lin JH. Health utility bias: A systematic review and meta-analytic evaluation. Med Decis Making. 2010; 30: 58-67.
16. Gafni A. The Standard Gamble Method—what is being measured and how it is interpreted. Health Serv Res. 1994; 29: 207-224.
17. Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care. 2008; 46: 357-365.
18. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul Health Metr. 2003; 1:1-12.
19. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics 2008; 26: 661-77.
20. Krabbe PFM, Devlin NJ, Stolk EA, Shah KK, Oppe M, van Hout B, Quik EH, Pickard AS, Xie F. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014; 52(11): 935-943.

CHAPTER 1

Head-to-head comparison of EQ-5D-3L and EQ-5D-5L health values

Selivanova A, Buskens E, Krabbe PFM. Head-to-head comparison of EQ-5D-3L and EQ-5D-5L health values. PharmacoEconomics 2018; 36(3): 715-725.

ABSTRACT

Background

The EQ-5D is a widely used preference-based instrument to measure HRQoL. Some methodological drawbacks of its three-level version (EQ-5D-3L) prompted development of a new format (EQ-5D-5L). There is no clear evidence that the new format outperforms the standard version.

Objective

To make a head-to-head comparison of the EQ-5D-3L and EQ-5D-5L in a discrete choice model setting giving special attention to the consistency and logical ordering of coefficients for the attribute levels and to the differences in health-state values.

Methods

Using efficient designs, 240 pairs of EQ-5D-3L and 240 pairs of EQ-5D-5L health states were generated in a pairwise choice format. The study included 3,698 Dutch general population respondents, analyzed their responses using a conditional logit model, and compared the values elicited by EQ-5D-3L and EQ-5D-5L for different health states.

Results

No inconsistencies or illogical ordering of level coefficients were observed in either version. The proportion of severe health states with low values was higher in the EQ-5D-5L than in the EQ-5D-3L, and the proportion of mild/moderate states was lower in the EQ-5D-5L than in the EQ-5D-3L. Moreover, differences were observed in the relative weights of the attributes.

Conclusion

Overall distribution of health state values derived from a large representative sample using the same measurement framework for both versions showed differences between the EQ-5D-3L and EQ-5D-5L. However, even small differences in the phrasing (language) of the descriptive system or in the valuation protocol can produce differences in values between these two versions.

1. INTRODUCTION

Generic preference-based measures of health-related quality of life (HRQoL) are frequently used to assess the impact of treatment or clinical pathways and to monitor population health [1-3]. Typically, preference-based measurement frameworks incorporate various independent attributes (notated for domains/dimensions) that jointly represent the notion of HRQoL. The levels of these attributes are weighted to indicate the relative importance attributed to them by the respondents (expressed preferences). Weighted attribute levels are subsequently aggregated into a single number reflecting the quality or value of a health state [4]. To obtain such values, several instruments (e.g., EQ-5D, HUI-3, SF-6D, AQol) have been developed within a preference-based measurement framework.

The EuroQol Group (www.euroqol.org) developed the EQ-5D, a relatively simple instrument that has been widely used [5-9]. It comprises five health attributes in the descriptive system (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) and a 20-cm visual analogue scale (VAS). In the standard version (EQ-5D-3L) each of the attributes can take on three levels [10]. A considerable body of literature corroborates the sustainability of the instrument [11-15]. However, attention has been drawn to its limited sensitivity regarding small or moderate changes in patients' health states [16-19] and its considerable ceiling effects (i.e., almost no differentiation between mild health states), prompting an update of the instrument [20-23]. In the new version, the EQ-5D-5L, the number of levels used to classify health states increased from three to five. Testing its descriptive system performance in terms of its discriminatory power and sensitivity revealed a lower ceiling effect and a higher sensitivity [13, 19, 23-25]. Additionally, several studies noted that subtle differences in the phrasing of levels 4 (severe problems) and 5 (extreme problems) caused inconsistencies in elicited health-state values [26-27].

Besides increasing the number of levels from three to five, the protocol to derive valuations was changed. For the EQ-5D-3L valuation protocol, originally the time trade-off (TTO) was chosen from among all possible health valuation techniques (standard gamble, time trade-off, rating/visual analogue scale, person trade-off, and magnitude estimation). However, various shortcomings of this technique were identified [28-31], which encouraged the EuroQol Group to experiment with other methods, such as choice-based modeling. Choice models are grounded in modern measurement theory and are consistent with the random utility model in economic theory [32]. The applicability of choice models for health-state evaluations has been proposed and tested elsewhere [4, 33-35].

The association between the descriptive systems for the three-level and the five-level versions of the EQ-5D has been investigated extensively. Far less is known about the distribution of the values and the underlying weights for the levels of the attributes for both EQ-5D versions, which motivated the present study. This paper presents a discrete choice study and head-to-head comparison of the EQ-5D-3L and EQ-5D-5L with an

emphasis on the consistency and logical ordering of attribute levels and the distributions of the estimated values.

2. METHODS

2.1 Sample

Overall, 4,036 persons participated in a self-completed computer-based assessment by SSI (Survey Sampling International, Rotterdam, Netherlands). The sample is representative of age and gender for the general Dutch population based on the SSI panel of working age 18-65 and was recruited in September-October 2016. Clear instructions were given to all participants, and those who fully completed the survey received a small financial compensation from SSI. The rewards were defined by the company's (SSI) internal agreements individually with the groups of respondents. Each one was randomly assigned to one of the 30 blocks of the survey. No limits on time for completion were imposed.

2.2 Discrete choice

Discrete choice (DC) modeling is a widely used technique to elicit personal and societal preferences in health-valuation studies [36]. The statistical literature classifies it within the modern framework of probabilistic discrete choice models that are consistent with economic theory (i.e., the random utility model) [32, 37-38]. All DC models establish the relative merit of one phenomenon based on its relative attractiveness. This technique requires participants to make choices among two or more presented scenarios (choice tasks) described by the means of specific attributes with certain levels.

2.3 Experimental design and selection of health states

The EQ-5D-3L contains five attributes with three levels each, yielding $3^5 = 243$ possible health states. Health states were presented in pairs for comparison in the DC task. Thus, the number of potential pairs to be compared becomes 29,403. For EQ-5D-5L the number of possible health states increases to $5^5 = 3{,}125$, and the number of possible paired comparisons rises drastically to 4,881,250. Clearly, it is infeasible to present all possible pairs to the respondents, especially in the case of EQ-5D-5L. For both versions, therefore, health-state pairs had to be carefully selected to arrive at an informative set. Two important issues were taken into consideration in the selection: respondent fatigue and avoidance of dominance in the pairs.

The credibility of one's responses can be questionable when a person gets bored or fatigued, which could happen if the tasks are complex or numerous. Earlier studies suggested that up to 16 choice tasks are acceptable and do not affect the responses [31, 39-40]. We offered each respondent a set of 16 choice tasks and reduced their complexity through two-level overlap in the health-state descriptions for

both versions of the EQ-5D. Two-level overlap implies fixing two of the five attributes at the same level and varying the other three.

Dominance is a common difficulty in health-state valuation exercises since all attributes are ordered, and people always prefer fewer health problems to more. Dominant pairs do not offer additional information, yet they reduce design efficiency. Therefore, it was decided to remove all combinations where every attribute of one health state in a pair was worse or the same (or better or the same) than every attribute of the other health state.

In view of the above solutions for the issues of fatigue and dominance, an approach to health-state selection was developed along similar lines, as set forth below for the EQ-5D-3L and the EQ-5D-5L. The set of non-dominant pairs for EQ-5D-3L was selected out of all possible 29,403 pairs, arriving at 14,580 pairs. Likewise, in EQ-5D-5L the number of non-dominant pairs was reduced from 4,881,250 to 1,430,000 (Stata 14.0). Out of all non-dominant health-state pairs with two-level overlap, we decided to select 240 pairs, which is considered sufficient to estimate regression coefficients for EQ-5D-5L attribute levels. It was decided to select the same number of pairs for the EQ-5D-3L. Therefore, 240 pairs in EQ-5D-5L and 240 pairs in EQ-5D-3L format were selected, using an efficient design routine programmed in Ngene software (the mnl model, taking 500 Bayesian draws, Halton sequence, modified Fedorov algorithm). All selected pairs were divided into 30 blocks with 16 choice tasks each, whereby 15 blocks contained all 16 tasks in EQ-5D-3L, and 15 blocks contained 16 tasks in EQ-5D-5L. The design was based on an iterative procedure, where designs are compared by their D-error (measure of statistical efficiency). After numerous iterations, the designs were checked for their D-errors and for the level balance. Level balance makes sure that all levels of all attributes appear evenly frequent in the design. Perfectly even frequency of level balance can rarely be achieved; therefore, the fairly even distribution of levels was accepted. Finally, the design with the lowest D-error and better indicator of level balance was chosen. Efficient design in Ngene requires priors (approximations of the parameters), which were derived from an earlier EQ-5D-3L study [36] and from a multinational study of the EQ-5D-5L [4].

2.4 Response tasks

The response task included two health-state descriptions comprised of the five attributes of the EQ-5D. The respondents had to decide which of the two health-state descriptions they preferred. Half of the blocks contained health-state descriptions defined by three levels of EQ-5D-3L (no problems, some problems, extreme problems), and half of the blocks contained health-state descriptions defined by five levels of EQ-5D-5L (no problems, slight problems, moderate problems, severe problems, extreme problems). The respondents were randomly assigned to one of the blocks, meaning that each person completed 16 response tasks only in EQ-5D-3L format or (in the other block) only in EQ-5D-5L format.

2.5 Analysis

2.5.1 EQ-5D-3L and EQ-5D-5L values and value' distributions

The analysis of the data was performed using a discrete choice conditional logit model (asclogit, Stata 15.0), which yields parameter estimates presented as regression coefficients. The main-effects value function included 10 dummy variables for the EQ-5D-3L representing level 2 and 3 for each of the five attributes: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). The main-effects model for the EQ-5D-5L included 20 dummy variables representing level 2, 3, 4, and 5. The regression coefficients were checked for logical ordering and significance. In addition, we tested for the increments from one level to any other consecutive levels (post-hoc estimation, contrast, Stata SE 15.0) [41, 42].

Additionally, the values of all health states possible in EQ-5D-3L and EQ-5D-5L were calculated based on estimated coefficients. We used the original values derived with the choice model and rescaled them to the published results of the Dutch valuation studies for the 3L version and 5L version respectively [43, 44]. For the EQ-5D-3L the value range from the valuation studies was -0.33 to 1.0, while for the EQ-5D-5L the value range was -0.45 to 1.0. Finally, for both versions the distributions of estimated values were compared. Kernel density graph and graphs of frequency distributions were produced for the EQ-5D-3L and the EQ-5D-5L (Stata SE 15.0). For comparison of value ranges in both graphs for EQ-5D-3L and the EQ-5D-5L, we provided distributions displaying the unscaled values and the values scaled to the Dutch tariffs.

2.5.2 Comparison of differences in weights for health-state attributes

The overall weights of each of the five EQ-5D attributes were calculated using the coefficient range method: the range between the coefficients of the individual levels was calculated and then converted to a proportion.

$$W_{attribute(i)} = \frac{\max Ci - \min Ci}{\sum_j (\max Cj - \min Cj)} \quad (3),$$

where C_i represents the coefficients of the individual levels of attribute *i* and *j* the number of attributes.

3. RESULTS

3.1 Sample

In total, 4,036 respondents completed the survey. Out of this sample, 288 completed 16 choice tasks in less than two minutes, which was considered unrealistic and insufficient. In addition, responses of 50 individuals were deemed unreliable, given their pattern of choosing only the left (A) or only the right (B) alternative throughout the survey. Therefore,

the forms of 338 respondents were disregarded. Finally, the analysis included 1,824 respondents for the EQ-5D-3L and 1,874 for the EQ-5D-5L (Table 1). An overall Chi^2 test revealed significant differences between the samples completing EQ-5D-3L and EQ-5D-5L in terms of age groups: P-value =0.000.

3.2 Comparison of EQ-5D-3L and EQ-5D-5L coefficients and overall attribute weights

No inconsistencies or illogical ordering of level coefficients were observed for the 3L and 5L versions. The spread of regression coefficients within each attribute consistently followed the same patterns across attributes: levels 2 and 3 lowered the values slightly, levels 4 and 5 even more so in the EQ-5D-5L. Moreover, the incremental differences between consecutive levels of each dimension were checked for significance, whereby it was observed that the move from level 5 to level 4 of severity had smaller effect than move from level 4 to level 3. All parameters in both models were statistically significant (Table 2 and 3).

Self-care was generally assigned less weight than the other four attributes in the EQ-5D-3L and in EQ-5D-5L (Table 2). Moreover, level 3 problems with mobility (confined to bed) appeared to have the largest effect on the values in the EQ-5D-3L format. Overall, the attribute mobility in the EQ-5D-3L version was assigned the highest relative weight. Regarding the EQ-5D-5L version, the respondents were more concerned about anxiety/depression and pain/discomfort than about problems with other attributes. Regarding the EQ-5D-3L version, we noted that pain/discomfort had more relative weight than anxiety/depression, while the opposite was noted for EQ-5D-5L.

3.3 Comparison of EQ-5D-3L and EQ-5D-5L value distributions

The original unscaled values of both EQ-5D versions were anchored to the values of the best and worst health states derived from the Dutch valuation studies [43, 44], and plotted as the frequency distribution of estimated values for 243 health states in the EQ-5D-3L and 3,125 health states in the EQ-5D-5L (Figure 1).

The graph demonstrates that the distributions of values elicited with unscaled coefficients are similar to the distribution of the rescaled values, because only the scale is changed, not the distribution of the values. These graphs and kernel density graph (Figure 2) demonstrate that EQ-5D-5L has more health states than EQ-5D-3L on the region with severe health states and fewer states on the region with milder states.

Table 1. Respondents' characteristics

Characteristics	EQ-5D-3L (N=1,824)	EQ-5D-5L (N=1,874)
Male, N (%)	797 (44)	876 (47)
Age, mean (SD)	45.5 (14.3)	51.2 (13.4)
Age group, N (%)		
18-24	101 (13)	74 (8)
25-34	90 (11)	47 (5)
35-44	134 (17)	95 (11)
45-54	214 (27)	172 (20)
Over 55	258 (32)	488 (56)
Female, N (%)	1,027 (56)	998 (53)
Age, mean(SD)	42.8 (13.8)	44.9 (15.1)
Age group, N (%)		
18-24	145 (14)	179 (18)
25-34	175 (17)	108 (11)
35-44	192 (19)	121 (12)
45-54	276 (27)	224 (22)
Over 55	239 (23)	366 (37)
Diseases, N (%)		
No diseases	701 (38)	705 (33)
Neck- and back pain	440 (24)	459 (25)
Pain (abdomen, migraine, chronic, etc.)	231 (13)	208 (11)
Sleep problems	258 (141)	281 (15)
Fatigue	337 (19)	360 (19)
Diabetes	132 (7)	163 (9)
Heart disease	94 (5)	140 (7)
Hearing or vision loss	149 (8)	182 (10)
Asthma/COPD	177 (10)	163 (9)
Eczema	126 (7)	145 (8)
Mental health problems	171 (9)	179 (10)
Stroke	16 (1)	37 (2)
Rheumatism (osteoarthritis, arthritis)	186 (10)	195 (10)
Cancer	27 (2)	46 (2)
Epilepsy	20 (1)	14 (0.5)
Lung disease	38 (2)	37 (2)
Gastrointestinal disease	63 (4)	64 (3)

Table 2. Regression coefficients for the EQ-5D-3L and EQ-5D-5L based on discrete choice model

EQ-5D-3L (the five attributes with their overall weights)	β (SE)	EQ-5D-5L (the five attributes with their overall weights)	β (SE)
Mobility (0.248)		Mobility (0.172)	
No problems (level 1)	reference	No problems (level 1)	reference
Some problems (level 2)	-0.323 (0.02)	Slight problems (level 2)	-0.138 (0.04)
Confined to bed (level 3)	-1.550 (0.03)	Moderate problems (level 3)	-0.290 (0.03)
		Severe problems (level 4)	-0.968 (0.04)
		Unable to (level 5)	-1.267 (0.04)
Self-care (0.146)		Self-care (0.156)	
No problems (level1)	reference	No problems (level 1)	reference
Some problems (level 2)	-0.318 (0.02)	Slight problems (level 2)	-0.098 (0.04)
Unable to (level 3)	-1.044 (0.03)	Moderate problems (level 3)	-0.297 (0.03)
		Severe problems (level 4)	-0.938 (0.04)
		Unable to (level 5)	-1.123 (0.04)
Usual activities (0.178)		Usual activities (0.175)	
No problems (level 1)	reference	No problems (level 1)	reference
Some problems (level 2)	-0.172 (0.02)	Slight problems (level 2)	-0.150 (0.04)
Unable to (level 3)	-1.055 (0.03)	Moderate problems (level 3)	-0.228 (0.03)
		Severe problems (level 4)	-0.969 (0.03)
		Unable to (level 5)	-1.302 (0.04)
Pain/discomfort (0.237)		Pain/discomfort (0.237)	
None (level 1)	reference	None (level 1)	reference
Moderate (level 2)	-0.247 (0.02)	Slight* (level 2)	-0.076 (0.04)
Extreme (level 3)	-1.423 (0.03)	Moderate (level 3)	-0.262 (0.04)
		Severe (level 4)	-1.150 (0.04)
		Extreme (level 5)	-1.636 (0.04)
Anxiety/depression (0.191)		Anxiety/depression (0.259)	
None (level 1)	reference	None (level 1)	reference
Moderate (level 2)	-0.379 (0.03)	Slight (level 2)	-0.253 (0.04)
Extreme (level 3)	-1.324 (0.03)	Moderate (level 3)	-0.543 (0.04)
		Severe (level 4)	-1.347 (0.04)
		Extreme (level 5)	-1.957 (0.04)
Log likelihood	-16979.542	Log likelihood	-16477.634
Wald chi2	4874.59	Wald chi2	5988.72

All variables were statistically significant at 99% confidence level, P-value < 0.01, except * P-value=0.037

Table 3. Estimations for the EQ-5D-3L and EQ-5D-5L increments for consecutive levels

EQ-5D-3L	β (SE)	EQ-5D-5L	β (SE)
Mobility		Mobility	
Some → no problems (level 2→level 1)	0.323 (0.02)	Slight → no problems (level 2→level 1)	0.138 (0.03)
Confined to bed → some problems (level 3→level 2)	1.227 (0.03)	Moderate → slight problems (level 3→level 2)	0.152 (0.03)
		Severe → moderate problems (level 4→level 3)	0.678 (0.03)
		Unable → severe problems (level 5→level 4)	0.298 (0.03)
Self-care		Self-care	
Some → no problems (level 2 →level 1)	0.318 (0.02)	Slight → no problems (level 2→level 1)	0.098 (0.04)
Unable → some problems (level 3→level 2)	0.726 (0.02)	Moderate → slight problems (level 3 →level 2)	0.199 (0.04)
		Severe → moderate problems (level 4→level 3)	0.641 (0.04)
		Unable → severe problems (level 5→level 4)	0.185 (0.04)
Usual activities		Usual activities	
Some → no problems (level 2→level 1)	0.172 (0.02)	Slight → no problems (level 2→level 1)	0.150 (0.04)
Unable → some problems (level 3→level 2)	0.884 (0.03)	Moderate → slight problems* (level 3→level 2)	0.079 (0.04
		Severe → moderate problems (level 4→level 3)	0.741 (0.04)
		Unable → severe problems (level 5→ level 4)	0.333 (0.04)
Pain/discomfort		Pain/discomfort	
Moderate → none (level 2→level 1)	0.247 (0.02)	Slight → none* (level 2→level 1)	0.076 (0.04)
Extreme → moderate (level 3→level 2)	1.176 (0.03)	Moderate → slight (level 3→level 2)	0.186 (0.04)
		Severe → moderate (level 4→level 3)	0.888 (0.04)
		Extreme → severe (level 5→level 4)	0.486 (0.03)
Anxiety/depression		Anxiety/depression	
Moderate → none (level 2→level 1)	0.379 (0.03)	Slight → none (level 2→level 1)	0.253 (0.04)
Extreme → moderate (level 3 →level 2)	0.945 (0.02)	Moderate → slight (level 3→level 2)	0.289 (0.03)
		Severe → moderate (level 4→level 3)	0.804 (0.03)
		Extreme → severe (level 5→level 4)	0.610 (0.04)

All variables were statistically significant at 99% confidence level, P-value < 0.01, except * P-value<0.05

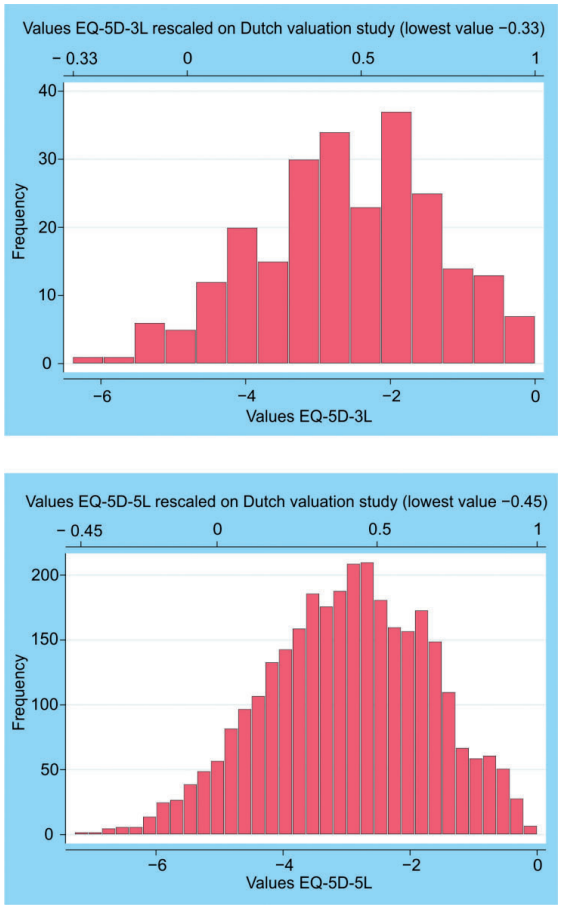

Fig. 1 Frequency distribution of (a) all 243 EQ-5D-3L health-state values and rescaled values; (b) all 3125 EQ-5D-5L health-state values and rescaled values

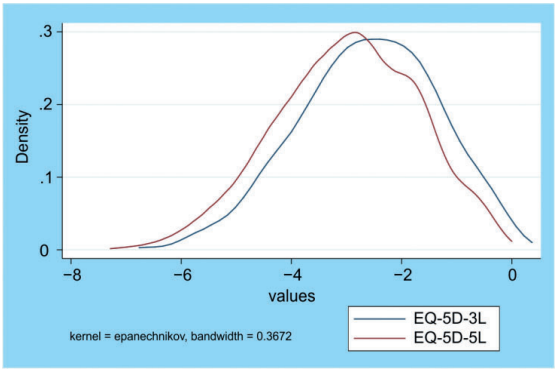

Fig. 2 Kernel density plot for EQ-5D-3L and EQ-5D-5Lvalues

4. DISCUSSION

4.1 Overall discussion and literature review

This study contributes to the body of literature comparing the standard EQ-5D-3L and the new EQ-5D-5L. Here, the focus is on the logical ordering and differences in distributions of values for health states in these two versions. The health-state values were elicited from a sample of the general population applying a conventional discrete choice approach. According to several earlier studies, the differences in EQ-5D-5L levels are subtle and may be hard to distinguish, which might have caused inconsistencies for some language versions (English) in the upper or lower levels of health attributes [26-27, 44]. Eventually, such inconsistency would affect the validity of the estimated values. In the current Dutch study, we found that all coefficients for both versions of EQ-5D were logically ordered.

However, the results demonstrated that the overall weights for the attributes are different in the two EQ-5D versions. In the EQ-5D-5L, the highest weight was attributed to anxiety/depression followed by pain/discomfort; in the EQ-5D-3L, the highest weight was attributed to mobility. Larger weights of an attribute has larger effects on a health-state value: the negative changes in the levels of the most important attributes could overweigh the positive changes in the levels of the less important attributes resulting into lower values.

Mobility, especially level 3 (confined to bed) had the most significant impact in the EQ-5D-3L. It is clear that 'confined to bed' has a different phrasing format for level 3 than it has in the other attributes. In later versions of the EQ-5D, namely the version for youth (EQ-5D-Y) and the EQ-5D-5L, the formulation of the worst levels was changed into 'unable to walk' [19,44-47]. In the EQ-5D-5L version, with the most severe level formulated as 'unable to', the effect of mobility on the health-state values declined. Changing the phrasing from 'confined to bed' to 'unable to walk' is likely responsible for the shift in the level of importance. 'Confined to bed' seems to imply isolation and dependence, while 'unable to walk' may be interpreted as a less serious limitation.

A large multinational study based on discrete choice modeling for the EQ-5D-5L [4] showed greater importance assigned to pain/discomfort and anxiety/depression attributes for the Dutch population, while for the US population the attribute mobility had the greater importance. The Dutch valuation study for the EQ-5D-5L confirmed that the greatest importance was assigned to pain/discomfort and anxiety/depression [44]. Mulhern et al. [48], in their study using discrete choice modeling, observed that the attribute pain/discomfort also showed the largest effect.

Overall, we observed differences in the health state distributions for severe and mild/moderate states derived from the EQ-5D-3L and EQ-5D-5L. Our findings are not in line with those of Mulhern [49], who observed the opposite. However, it may be attributed to the fact that for the EQ-5D-3L UK value set has larger range of values than the EQ-5D-5L UK value set. In addition, the samples analyzed in that study were recruited differently (UK, England) and different valuation methods were used (time

trade-off, visual analogue scale). Overall, the distributions of health states in the current study showed somewhat lower proportion of severe states in the EQ-5D-5L than in the EQ-5D-3L. These findings are not in line with the findings published in the Dutch tariff [44], demonstrating the values for all attainable health states to be higher in 3L version for the severe health states and higher for 5L version for moderate and mild health states (on the value range 0.35-0.75). Again, such discrepancy may be caused by differences in conceptual and valuation approaches used. The current study is based only on DC estimations, while the Dutch tariff is based on the composite TTO and tasks for valuing worse-than-death states were included. In the Dutch tariff study DC results were used to identify the appropriate TTO modeling techniques, but not to estimate health state values.

The report by NICE [50] suggested that the 5L instrument showed higher mean utility scores than the 3L, meaning that the improvements in health are slightly less in the 5L than in the 3L, which results into interventions considered as less cost-effective if based on the 5L. This may lead policymakers to give due consideration to the choice of a version: EQ-5D-5L may produce smaller benefits of innovations for severe health states, according to our study, which may discourage end-users from using this version. These findings raise challenges about the choice of the EQ-5D version to be used: for particular interventions end-users are likely to prefer the EQ-5D-3L indicating higher benefits of interventions. However, the studies included in the NICE document are not based on valuations. In fact, the analysis underlying that document used self-reported health assessments scored according to the EQ-5D descriptive system. Therefore, the comparison between the current study and the study of NICE should be taken with caution.

4.2 Limitations

It is worth mentioning the following limitation of our study: there is a difference in the age groups proportions of the two samples. We tried to reach the comparability of the representativeness and sample sizes for 3L and 5L versions, however, significant age differences were observed according to the Chi^2 test. One might argue that such differences would bias the estimated results. However, an additional analysis with inclusion of age groups as a separate predictor into the choice model did not reveal any statistically significant effect of age on the estimated coefficients.

By their nature, health-state values derived with choice models cannot be interpreted as absolute (cardinal) numbers due to two reasons. First, the best health state (full health) is dominant and cannot be used in the choice model as anchor. Second, the location of death is unknown since a 'death' option was not included. Consequently, DC models position health states on a scale between the best and the worst health states. Therefore, one of the main problems with choice models is normalizing its scale to a death-full health (0.0 – 1.0) scale. To solve this problem, a task extension or additional

tasks should be included on the design, like death questions, duration on the health states or an accompanied TTO task. We did not use either of these techniques. Instead, we used the published Dutch valuation studies [43, 44] as anchor for the values elicited with the discrete choice model. By doing so, the rescaling limitation remains but anchor points are based in current evidence.

Recent studies using different valuation frameworks for QALYs calculations showed smaller differences between the same health states in the EQ-5D-5L version in comparison with the original EQ-5D-3L, which raised concerns among end-users (e.g., pharmaceutical companies) [44, 49, 50]. In a recent UK study estimating a value function for the EQ-5D-5L, the composite TTO was introduced as a new valuation technique. That innovation is a derivate of the conventional TTO based on a combination of lead-time TTO [51] and standard TTO as used in the 3L. This UK study applied a rescaling for the states 'worse than death' (negative utilities) that differs from the rescaling used in the original EQ-5D-3L [1]. On top of that, the UK study [52] analyzed TTO responses and DC responses together in a hybrid model incorporating several other analytical procedures (e.g., censoring, additional parameter for heterogeneity of respondents, forcing consistency in levels of attributes) [53]. Moreover, the authors of the Dutch tariff [44] admitted that the similarities between the EQ-5D-3L and EQ-5D-5L are not necessarily expected due to differences in phrasing and valuation methods used. Therefore, the divergence between the 3L and 5L version, if based on the official EuroQol protocol, is likely to be a combined effect of the differences in the way individuals respond to the changed descriptive system and because a totally new and different valuation framework has been introduced [54]. The present study did not use a time trade-off technique. Instead we used DC for both versions of EQ-5D, which resulted in certain differences in the weights and overall distributions of the EQ-5D-3L and the EQ-5D-5L health-state values. Values derived with DC seem to be more robust and less effected by possible framing effects, as the judgmental DC task is more straightforward and simple than the TTO variants. However, it needs to be stated that the design strategy of selecting equal amount of DC pairs for both versions may have had an impact on the estimated values. Specifically, selecting 240 DC pairs for the EQ-5D-3L would enable broader coverage of the health-states than selecting 240 pairs for the EQ-5D-5L, since the EQ-5D-5L comprise more health-states. Consequently, such design setting would result in more precise estimates for the EQ-5D-3L than for the EQ-5D-5L. However, based on earlier studies [4, 44, 48], having 240 pairs for the EQ-5D-5L is highly sufficient to get precise estimates. Moreover, the standard deviations of the coefficients, which reflect precision of an estimated coefficient, showed that the difference is minor (maximum SD in the model for EQ-5D-3L is 0.3, while maximum SD in the model for EQ-5D-5L is 0.4).

4.3 Strengths

The present study has several strengths. First, a large representative sample of the Dutch general population has been achieved. Second, it used the same valuation method (discrete choice) and the same statistical analysis for both EQ-5D versions. Third, an efficient design was applied to maximize the precision of estimated regression coefficients, while respondent fatigue was prevented by applying two-level overlap. Overall, this is the first head-to-head discrete choice study to compare health-state values derived from EQ-5D-3L and EQ-5D-5L using large samples.

4.4 Conclusion

In conclusion, the distributions of health states suggested that proportion of severe health states with low values in the EQ-5D-5L was slightly higher than in the EQ-5D-3L, and the proportion of mild/moderate states was lower in the EQ-5D-5L than in the EQ-5D-3L.

Additionally, the overall weights of the attributes in the EQ-5D-3L and the EQ-5D-5L are different. We suggest that even small differences in the phrasing of the descriptive system or in the valuation protocol may affect individual responses and thereby the elicited values. Finally, it needs to be emphasized that the applied valuation framework in combination with particular statistical models used to estimate the weights for the attributes and their levels, may explain the substantial discrepancies between the 3L and 5L observed in earlier studies.

5. REFERENCES

1. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997; 35(11):1095-1108.
2. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M. Multiattribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. Med Care. 2002; 40(2): 113–128.
3. Hamming JF, De Vries J. Measuring quality of life. Br J Surg. 2007; 94: 923–924.
4. Krabbe PFM, Devlin NJ, Stolk EA, Shah KK, Oppe M, van Hout B, Quik EH, Pickard AS, Xie F. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014; 52(11): 935-943.
5. Hurst N, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in people with rheumatoid arthritis: validity, responsiveness and reliability of the EuroQoL (EQ-5D). Br. J. Rheumatol. 1997; 36: 551–559.
6. Rabin R, Charro de F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001; 33:337-343.
7. Russell RT, Feurer ID, Wisawatapnimit P, and Pinson CW. The validity of EQ-5D US preference weights in liver transplant candidates and recipients. Liver Transpl. 2009; 15: 88–95. doi:10.1002/lt.21648.
8. Xu R, Insinga RP, Golden W, Hu XH. EuroQol (EQ-5D) health utility scores for patients with migraine. Qual Life Res. 2011; 20(4): 601-608.
9. Devlin NJ & Brooks R. EQ-5D and the EuroQol Group: Past, Present and Future. Appl Health Econ Health Policy. 2017; 15: 127. https://doi.org/10.1007/s40258-017-0310-5
10. Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37 (1), 53–72.
11. Johnson JA, Coons SJ. Comparison of the EQ-5D and SF-12 in an adult US sample. Qual Life Res. 1998; 7:155–66.
12. Johnson JA, Pickard AS. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. Med Care. 2000; 38 (1), 115–121.
13. Pickard AS, De leon MC, Kohlmann T,Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007; 45: 259–63.
14. Dyer MT, Goldsmith KA, Sharples LS Buxton MJ. A review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes. 2010; 8:1–13.
15. Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. Diabet Med. 2011; 28:395–413.
16. Myers C, Wilks D. Comparison of Euroqol EQ-5D and SF-36 in patients with chronic fatigue syndrome. Qual Life Res. 1999; 8: 9. doi:10.1023/A:1026459027453.
17. Wu AW, Jacobson KL, Frick KD, Clark R, Revicki DA, Freedberg KA, Scott-Lennox J, Feinberg J. Validity and responsiveness of the EuroQol as a measure of health-related quality of life in people enrolled in an AIDS clinical trial. Qual Life Res. 2002; 11: 273–82.
18. Macran S, Weatherly H, Kind P. Measuring population health: a comparison of three generic health status measures. Med Care. 2013; 41(2): 218–231.
19. Janssen MF, Pickard AS, Golicki D,Gudex C, Niewada M, Scalone L, Swinburn P, Bussbach J. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013; 22: 1717–27.

20. Badia X, Herdman M, Kind P: The influence of ill-health experience on the valuation of health. Pharmacoecon. 1998, 13:687-696.
21. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D-3L and SF-6D across seven patient groups. J Health Econ. 2004; 13(9):873–84.
22. Sullivan PW, Lawrence WF Jr, Ghushchyan V. A national catalogue of preference-based scores for chronic conditions in the U.S. Med Care. 2005; 43: 736–49.
23. Scalone L, Ciampichini R, Fagiuoli S, Gardini I, Fusco F, Gaeta L, Del Prete A, Cesana G, Mantovani LG. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic disease. Qual Life Res. 2013; 22: 1707–16.
24. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value Health. 2008; 11: 275–84.
25. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011; 20(10): 1727–1736.
26. Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PFM. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best–worst scaling? Eur J Health Econ. 2014. doi: 10.1007/s10198-013-0474-3.
27. Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? Qual Life Res. 2015; 24(7): 1759–1765.
28. Van Osch SMC, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting Biases in Standard Gamble and Time Tradeoff Utilities. Med Decis Making. 2004; 511-517.
29. Van der Pol M, Roux L. Time preference bias in time trade-off. Eur J Health Econ. 2005; 107-11.
30. Doctor JN, Bleichrodt H, Lin JH. Health Utility Bias: A Systematic Review and Meta-Analytic Evaluation. Med Decis Making. 2010; 30: 58-67.
31. Viney R, Norman R, Brazier J, Cronin P, King MT, Ratcliffe J, Street D. An Australian choice experiment to value EQ-5D health states. J Health Econ. 2014; 23:729-742.
32. Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: Background, theories, assumptions and applications. Expert Rev Pharmacoecon Outcomes Res. 2013; 13(1): 93–108.
33. McKenzie L, Cairns J, Osman L. Symptom-based outcome measures for asthma: the use of discrete choice methods to assess patient preferences. Health Policy. 2001; 57:193–204.
34. Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M. Using dce and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. J Health Econ. 2009; 18: 1261–1276.
35. Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate societal health state utility values. J Health Econ. 2012; 31: 306–318.
36. Stolk EA, Oppe M, Scalone L, Krabbe PFM. 2010. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. Value Health. 2010; 13, 1005-1013.
37. Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care. 2008; 46(4), 357-365.
38. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law. 2009; 4, 527-546.

39. Coast J, Flynn TN, Salisbury C, Louviere J, Peters TJ. Maximising responses to discrete choice experiments: A randomised trial. Appl Health Econ Health Policy. 2006; 5: 249–260.
40. Hall J, Fiebig DG, King MT, Hossain I, Louviere JJ. What influences participation in genetic carrier testing? Results from a discrete choice experiment. J Health Econ. 2006; 25: 520–537.
41. Ramos-Goñi JM, Craig BM, Oppe M, Ramallo-Fariña Y, Pinto-Prades JL, Luo N, Rivero-Arias O. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. Value Health. 2017; https://doi.org/10.1016/j.jval.2017.10.02
42. Finn JD. The selection of contrast. In: Holt, Rinehart and Winston. A General Model for Multivariate Analysis. New York, US; 1974.
43. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006; 15(10):1121-32.
44. Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016; 19 (4): 343-352.
45. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012; 15: 708 – 715.
46. Wang P, Luo N, Tai ES, Thumboo J. The EQ-5D-5L is more discriminative than the EQ-5D-3L in patients with diabetes in Singapore. Value Health. 2016; 9C: 57 – 62.
47. Burström K, Bartonek A, Broström EW, Sun S, Egmar A-C. EQ-5D-Y as a health-related quality of life measure in children and adolescents with functional disability in Sweden: testing feasibility and validity. Acta Pædiatrica. 2014; 103: 426–435.
48. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using discrete choice experiments with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. Med Dec Making. 2016; 1-13.
49. Mulhern B., Feng Y., Shah K., van Hout B., Janssen B., Herdman M., Devlin N. Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets. A report by the Centre for Health Economics Research and Evaluation 2017. https://www.ohe.org/publications/comparing-uk-eq-5d-3l-and-english-eq-5d-5l-value-sets. Accessed 4 July 2017.
50. Wailoo A, Alava MH, Grimm S, Pudney S, Gomes M, Sadique Z, Meads D, O'Dwyer J, Barton G, Irvine L. Comparing the EQ-5D-3L and 5L versions. What are the implications for cost effectiveness estimates? Report by the decision support unit 2017. http://scharr.dept.shef.ac.uk/nicedsu/wp-content/uploads/sites/7/2017/05/DSU_3L-to-5L-FINAL.pdf. Accessed 17 Aug 2017.
51. Janssen BMF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013; 14 (1):5-13.
52. Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing Health-Related Quality of Life: An EQ-5D-5L value set for England. Health Econ. 2017; 1-16.
53. Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: An application to English data. Health Econ. 2017; 1-16.
54. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014; 17(4): 445–453.

CHAPTER 2

Does inclusion of interactions result in higher precision of estimated health-state values?

Accepted for publication in Value in Health

ABSTRACT

Objective

Most preference-based instruments producing overall values for health states are devised on the simplifying assumption that the overall effect of distinct HRQoL domains (attributes) of the instrument equals the sum of its attributes. However, health aspects are often interrelated and depend on each other. Therefore, the objective is to investigate whether inclusion of second-order interactions in the EQ-5D-3L value function would result in better fit and lead to different health-state values than a model with main effects only.

Methods

Using an efficient design, 400 pairs of EQ-5D-3L health states were generated in a pair-wise choice format. We analyzed responses of 4,000 persons from the general population using a conditional logit model, and we tested goodness-of-fit using pseudo R^2, AIC, differences in log-likelihood, and likelihood ratio.

Results

The interactions model showed systematically lower values than the main effects model. Inclusion of interactions resulted only in a slightly better model fit. Interactions comprising mobility and self-care were the most salient.

Conclusion

For the EQ-5D-3L, a value function based on interactions produces systematically lower values than a main-effects model, meaning that the effects of two or more health problems combined is stronger than the sum of the individual main effects.

1. INTRODUCTION

A construct commonly used in health outcomes measurement is health-related quality of life (HRQoL), a subjective measure of perceived health status consisting of physical, mental, and social domains [1, 2]. One common framework to measure HRQoL is by preference-based measurement methods. Instead of measuring the level of reported complaints (i.e., their frequency and intensity) for distinct health domains, these methods express the quality of a patient's health condition. Preference-based measures differ from other approaches to measure health condition in that they explicitly incorporate weights reflecting the importance attached to a set of specific health domains (technical term: attributes) that each capture a specific health aspect. The measures produced by these methods are expressed in a single metric number, which we here refer to as 'value'. The core of a preference-based measurement framework consists of a response task comparing at least two objects (in the present case health condition) and to express which object is preferred (is better). Often the structured description of a health condition is referred to as a health state: a small set of attributes each with a limited number of levels of severity. The respondents do not score the attributes one by one but consider the whole set of health attributes, which requires reading and mentally processing all of the attributes in the set simultaneous [3]. The response task is to compare complete attribute sets, differing according to levels of severity, or comparing sets with a specified health outcome (e.g. immediate dead or living in full health for a specified number of years). By these comparisons a preference for one of the combinations of health states or health outcomes is evoked. There are several techniques allowing health-state evaluation within preference-based framework, but in the present study we chose the more recently introduced method of discrete choice modelling. Discrete choice modelling is widely used to elicit personal and societal preferences in health-valuation studies [4]. Discrete choice is considered a relatively easy task for the respondents since it mimics individual everyday choices: 'Which of the available options is more preferable?' (Figure 1).

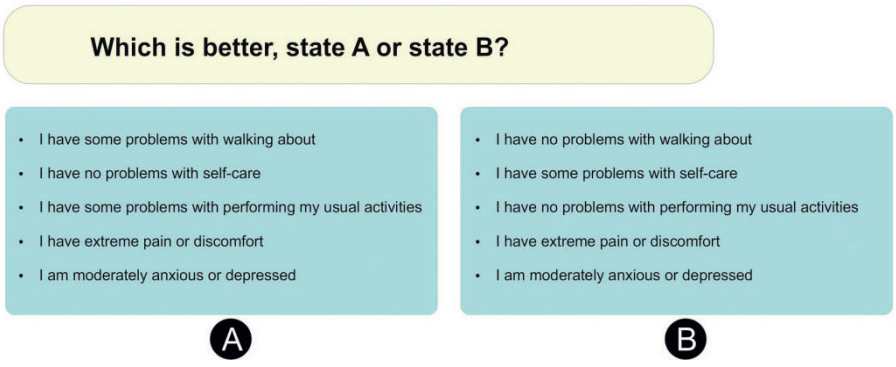

Fig. 1 Example of a discrete choice task for the EQ-5D-3L

The total number of states to be valued is determined by the possible level combinations of the classification. If there are few states, it may even be feasible to value them all. If there are many, a well-chosen subset (constructed in such a way to maximize information derived from a limited set of states out of all possible states) can be valued empirically, and the values for the remaining states can be estimated (usually by regression modeling). The values produced by these preference-based systems can be implemented in health-outcomes research, disease-modeling studies, economic evaluations to compare different healthcare interventions, and the planning and monitoring of health programs. The most common preference-based instruments (e.g., SF-6D or 15D) were developed using value functions comprising only main effects and ignoring the interactions between health attributes [5, 6]. Main-effect functions rely on the simplifying assumption that the overall effect of all HRQoL attributes equals the sum of the attribute levels included in the function. Interactions play a role when the overall effect of two separate attributes is significantly more (or less) than their individual effects (e.g., reduction in perceived health status may intensify if two different health problems interact). However, health attributes are often related and considered to depend on each other. Only for the HUI (Health Utility Index, 7 attributes with 5 or 6 levels per attribute) and AQoL (Assessment of Quality of Life, has versions with 4, 6, 7 or 8 attributes with multiple attributes) were interactions taken into account. However, by using a multiplicative model the interactions among all attributes were forced to be the same [2, 7]. Other explorative studies [4, 8] demonstrated that the effect of health-state attributes is not simply additive and that interactions may be important. However, this assumption has not yet been tested thoroughly for preference-based instruments [9-11].

Using the EQ-5D-3L instrument, this study investigates whether the inclusion of interaction terms leads to different estimated values for health states, and whether a model with interactions has better fit than a main-effects model.

2. METHODS

2.1 EQ-5D-3L instrument

The EQ-5D instrument was developed by the EuroQol Group (www.euroqol.org) as a relatively simple generic preference-based instrument that could be used in clinical studies and would provide values of health states for use in economic evaluations [12]. The EQ-5D-3L descriptive system comprises five attributes: mobility (MO); self-care (SC); usual activities (UA); pain/discomfort (PD); and anxiety/depression (AD). Each attribute has three levels: no problems, some problems, and severe problems. EQ-5D-3L health states are defined by selecting one level from each attribute, with 11111 denoting perfect health (no problems in any attributes) and 33333 the worst possible health state (severe problems in all attributes). While developing the EQ-5D, researchers were experimenting with various valuation techniques and considered discrete choice modelling as a promising alternative

to the conventional valuation techniques (time trade-off, standard gamble, visual analogue scale). However, the produced health values were based on value functions comprising only main effects and were produced by other than discrete choice methodology [13-14]. Simple additive value functions comprising main effects assumed that each of the five attributes was independent of others, ignoring the effects of any other attribute or their interactions [15].

2.2 Discrete choice modeling

Discrete choice (DC) modeling is a widely used technique to elicit personal and societal preferences in health-valuation studies [4]. The statistical literature classifies it within the modern framework of probabilistic discrete choice models that are consistent with economic theory (i.e., the random utility model) [16-19]. Discrete choice modeling is based on probabilistic statistical routines (logit or probit regression models), and are used to establish the relative merit of one phenomenon relative to others [20-21]. Such choice models allow estimating the relative importance of health-state specific attributes with certain levels, and overall values for health states with different combinations of attribute levels.

2.3 Health states selection

The EQ-5D-3L contains five attributes with three levels each, yielding $3^5 = 243$ possible health states. Presenting health states as paired comparisons in the discrete choice task (two health states being assessed together) increases this number to 29,403 possible combinations. The evaluation of all possible combinations is known as a full factorial design, which allows the researcher to estimate all main effects and all possible interaction effects. In practice this design is rarely used, as it is considered tedious and/or cost-prohibitive [22]. Another practical deterrent is that it usually entails very large sample sizes, a requirement that cannot always be met. These conditions explain why full factorial designs are almost never used for the valuation of health states, and even rarely in the field of marketing.

Fractional designs were developed to facilitate the careful selection of a subset of choice tasks out of all possible combinations. A carefully selected subset should be sufficient to reveal all important information for the investigated issue (in our case attributes with their different levels in each of the two health-state descriptions), while using only part of experimental efforts necessary for the full factorial design [23]. A fractional design was applied in the present study. The first step was to determine how many pairs of health state pairs to include in the design. This number should be sufficient for estimating all main effects and all second-order interaction effects for the EQ-5D-3L. In discrete choice models, the minimum criterion implies that the number of choice tasks is defined by the number of parameters. Specifically, the minimum amount exceeds by one the number of parameters needed to estimate in the model. The attribute

levels used for the present study are categorical variables, which are represented by dummy variables: MO1 (no problems with mobility), MO2 (some problems with mobility), MO3 (confined to bed), SC1 (no problems with self-care), SC2 (some problems with washing or dressing), SC3 (unable to wash or dress myself) , UA1 (no problems with usual activities), UA2 (some problems with usual activities), UA3 (unable to perform usual activities), PD1 (no pain/discomfort), PD2 (moderate pain/discomfort), PD3 (extreme pain/discomfort), AD1 (no anxiety/depression), AD2 (moderate anxiety/depression), and AD3 (extreme anxiety/depression). Effects coding was used in the design of the study, whereby level 3 was chosen as reference (omitted). Therefore, the main-effects model included eleven parameters for all non-omitted attributes at levels 1 and 2 (no problems and some problems), summing up to ten parameters to estimate Equation 1. Expressed as a formula, the model predicts latent values (V) of individuals choosing health states, where $\beta_{1\text{-}10}$ represents unknown regression coefficients, and (MO1, MO2, SC1, SC2 … AD2) are alternative-specific explanatory variables. In effects coding, the effects of the reference variable (level 3) can be derived as a negative summation of the effects of all non-omitted levels (level 1 and 2). For example, the effect of level 3 mobility is calculated as - (β_1MO1+β_2MO2).

$$V_s=\alpha+\beta_1 MO1+\beta_2 MO2+\beta_3 SC1+\beta_4 SC2+\beta_5 UA1+\beta_6 UA2+\beta_7 PD1+\beta_8 PD2+\beta_9 AD1+\beta_{10} AD2 \quad \text{(Eq. 1)}$$

The interaction model included the intercept, all main effects (ten parameters), and all second-order interactions between levels 1 and 2 (40 parameters) resulting in 51 parameters. This implies that at least 52 pairs of health states are needed to identify the model (Equation 2).

$$\begin{aligned}V_s=&\alpha+\beta_1 MO1+\beta_2 MO2+\beta_3 SC1+\beta_4 SC2+\beta_5 UA1+\beta_6 UA2+\beta_7 PD1+\beta_8 PD2+\beta_9 AD1+\beta_{10} AD2+\\&\beta_{11} MO1\times SC1+\beta_{12} MO1\times SC2+\beta_{13} MO2\times SC1+\beta_{14} MO2\times SC2+\beta_{15} MO1\times UA1+\beta_{16} MO1\times UA2+\\&\beta_{17} MO2\times UA1+\beta_{18} MO2\times UA2+\beta_{19} MO1\times PD1+\beta_{20} MO1\times PD2+\beta_{21} MO2\times PD1+\beta_{22} MO2\times PD2\varepsilon+\\&\beta_{23} MO1\times AD1+\beta_{24} MO1\times AD2+\beta_{25} MO2\times AD1+\beta_{26} MO2\times AD2+\beta_{27} SC1\times UA1+\beta_{28} SC1\times UA2+\\&\beta_{29} SC2\times UA1+\beta_{30} SC2\times UA2+\beta_{31} SC1\times PD1+\beta_{32} SC1\times PD2+\beta_{33} SC2\times PD1+\beta_{34} SC2\times PD2+\\&\beta_{35} SC1\times AD1+\beta_{36} SC1\times AD2+\beta_{37} SC2\times AD1+\beta_{38} SC2\times AD2+\beta_{39} UA1\times PD1+\beta_{40} UA1\times PD2+\\&\beta_{41} UA2\times PD1+\beta_{42} UA2\times PD2+\beta_{43} UA1\times AD1+\beta_{44} UA1\times AD2+\beta_{45} UA2\times AD1+\beta_{46} UA2\times AD2+\\&\beta_{47} PD1\times AD1+\beta_{48} PD1\times AD2+\beta_{49} PD2\times AD1+\beta_{50} PD2\times AD2 \quad \text{(Eq. 2)}\end{aligned}$$

After consideration of the number of choice tasks used in earlier studies [24-26] and the criteria for the number of choice tasks to include in the design, it was decided to increase the number of pairs to 400. That would allow for a wider range of estimated health states with various severity levels.

2.4 Experimental design

Interaction models are rarely applied because of their complexity due to the large amount of health state pairs to be judged by respondents. Judging large amounts of pairs by the same respondent can result in respondents' fatigue. To avoid this, researchers need to develop a design, optimal in terms of statistical and response efficiency, in which different blocks of pairs are offered to different set of respondents. In our study we used the following approach: the set of 400 health-state pairs was divided into 25 blocks with 16 choice tasks each. Earlier studies suggested that 16 choice tasks would be acceptable to the respondents and would not affect their responses [24, 27, 28]. Reliability may be questionable if the respondents are bored or fatigued. Burden can be caused either by task complexity or by having a large number of tasks to carry out. The complexity of the tasks was reduced by implementation of two-level overlap for the health states, and the number was limited to 16 choice tasks per respondent. The two-level overlap implies fixing two out of five attributes at the same level while the other three attributes can vary.

A common problem in health-state valuation exercises is dominance, since all attributes are ordered, and smaller health problems are always preferred to bigger ones. Dominant pairs do not offer additional information but instead reduce the design's statistical efficiency (variability of parameter estimates rises; standard errors are getting larger). Therefore, such combinations, where for one health state all the attributes were worse (or better) than those of its paired state, were removed from the candidate pairs for constructing the design. The set of possible pairs without dominant combinations and with two-level overlap was selected out of all possible 29,403 pairs. Out of the resulting set of 14,580 pairs, 400 health-state pairs were selected using an efficient design (Ngene software, the mnl model, taking 500 Bayesian draws, Halton sequence, modified Fedorov algorithm). An experimental design is called statistically efficient if the parameters are estimated with the least possible standard errors. Additional to statistical efficiency there is response efficiency. This means that respondents are offered tasks with reduced complexity to avoid attentional failures and failure in memory, thereby, getting more reliable responses. The design was constructed using an iterative procedure, whereby designs were compared in terms of their D-error, which is the measure of statistical efficiency we decided to use. D-errors were computed on the basis of expected values of the model parameters. Generation of an efficient design in Ngene requires prior distributions of the parameters, which were derived from a previous EQ-5D-3L study [4]. Because that study was not aimed at interaction estimations, only priors for main effects were set accordingly, and the priors for interactions were set to zeros.

2.5 Sample recruitment

According to the golden rule formulated by Johnson and Orme [29], $N>500c/(t\times a)$, where N (minimum sample size per a block of a survey), c (largest product of levels in interactions), t (number of tasks), and a (number of alternatives). In the model with

second-order interactions for EQ-5D-3L in the present study, c=9, t=16, a=2, number of blocks is 25, and calculated minimum sample size is N>500*9/32*25 =141*25=3,525, although more sophisticated calculation procedure may be found [30]. In discrete choice modeling, a total of 50-60 observations per response task would generally be considered sufficient. Based on this number of observations per choice set, the minimum sample size for 400 response tasks (400 pairs) is 1,500. The final sample for the present study consisted of 4,000 members of the Dutch general population of working age 18-65, representative on age and gender. The respondents were recruited using the panel of the marketing company SSI (Rotterdam, The Netherlands). Possible drop-outs or insufficient quality of responses, which could diminish the size of the sample eligible for the final analysis, were accounted for. Responses were assumed to be of insufficient quality when the completion time fell below two minutes, which was considered too short to perform 16 choice tasks carefully.

2.6 Analysis

The conditional logit routine was used to obtain coefficients of the EQ-5D-3L attribute levels from both models of interest: the main-effects model and the model with all second-order interactions (Stata 14.0). Since the research question of the current study focuses on overall values and not on heterogeneity among respondents, the basic conditional logit model was considered as sufficient [16]. In the latter model, the estimated coefficients represented the effects of attribute levels, and the interactions between the separate levels of one attribute versus the separate levels of another attribute. The overall significance of the ten second-order interactions (MO×SC, MO×UA, MO×PD, MO×AD, SC×UA, SC×PD, SC×AD, UA×PD, UA×AD, PD×AD) was not estimated on the basis of coefficients. Rather, it was tested on the basis of the likelihood ratio to conclude whether adding the interactions improved the model fit. The likelihood ratio was calculated for the model with all second-order interactions and the model without one specified interaction (i.e., MO × SC). If the P-value in the likelihood ratio test is low (below 0.05), the goodness-of-fit of the model with specified interactions is deemed significantly better than the goodness-of-fit without the specified interactions.

The goodness-of-fit for the model with main effects only and the model with interaction effects was investigated using pseudo R^2 and AIC. The higher pseudo R^2 and the lower AIC indicate better model fit. In addition, mean absolute error (MAE) and root-mean-square error (RMSE) were calculated to assess the accuracy of predictions of both models. MAE and RMSE present the differences between observed and predicted values from each model, therefore, reflecting the accuracy of models' predictions.

To demonstrate the differences between the estimates for the main-effects model and the interaction-effects model, predicted values of 243 EQ-5D-3L health states were plotted against each other (SigmaPlot 13.0). The value for the alternative in a choice task is modeled as the product of the health-state characteristics (severity of an attribute,

such as level 1 problems with mobility or level 2 problems with self-care) and health-state preference parameters (β). It needs to be noted, that in conditional logit model the constant term α was not shown since it does not vary across the alternatives. For instance, having parameter estimates for non-omitted levels 1 and 2 from conditional logit model, and calculating estimates for level 3 as the negative summation for the effects of all non-omitted levels (levels 1 and 2), we can calculate predicted value for the health state 23112 on the basis of the main-effects model (Equation 4) as follows:
U = βMO2 - (βSC1+ βSC2) + βUA1 +βPD1+βAD2=0.351-(0.488+0.084) +0.393+0.563+0.205 = 0.94. (Eq. 4)

For the interaction-effects model, the estimates for all 243 health states were calculated by summation of main-effects and interaction-effects coefficients of levels comprising the health state. Consider, for example, the calculation for health state 23112 (Equation 5):
U = βMO2 - (βSC1+ βSC2) + βUA1 +βPD1+βAD2-(βMO2×SC1+ βMO2×SC2) + βMO2×UA1 + βMO2×PD1 + βMO2×AD2 - (βSC1×UA1+βSC2×UA1) - (βSC1×PD1+βSC2×PD1) - (βSC1×AD2+βSC2×AD2) + βUA1×PD1 + βUA1×AD2 + βPD1×AD2 = 0.329-0.565+0.397+0.572+0.194-0.001+0.043-0.048+0.021-0.019-0.043-0.034+0.040-0.002-0.023=0.86. (Eq. 5)

The given calculations of values (Eq. 4, Eq.5) are based on unscaled model coefficients (i.e., values are not scaled from 0 to 1). To see whether the health-state values in the main-effects model differ from the health-state values in the model including second-order interactions, the values of all health states were rescaled from 0 (worst health state 33333) to 1 (best health state 11111) and then plotted.

3. RESULTS

3.1 Sample

The survey was completed by 4,000 respondents aged between 18 and 65. However, 309 respondents were removed from the analysis because their responses were deemed unreliable due to the short amount of time spent on the survey (less than two minutes). Before the analysis, 22 respondents were discarded due to the observed pattern of choosing only the left or only the right alternative. Ultimately, 3,669 respondents were included in the final analysis. The representative sample from the Dutch population was recruited in October 2016 (Table 1).

3. 2 Main- and interaction-effect models

In the main-effects model for EQ-5D-3L, all estimates are logically ordered and statistically significant at the 95% level (Table 2).

Table 1. Respondents' characteristics

Characteristics	Respondents N=3,669
Male, N (%)	1,645 (45)
Age, mean(SD)	46.0 (13.4)
Age group, N (%)	
18-24	145 (9)
25-34	219 (13)
35-44	316 (19)
45-54	426 (26)
Older than 55	539 (33)
Female, N (%)	2,024 (55)
Age, mean(SD)	42.5 (13.8)
Age group, N (%)	
18-24	313 (15)
25-34	329 (16)
35-44	394 (20)
45-54	529 (26)
Older than 55	459 (23)

Table 2. Parameter estimates for main-effects model based on discrete choice (DC) data, effects coding

	Main-effects estimates	
	β (SE)	P-value
MO1	0.618 (0.01)	0.000
MO2	0.351 (0.01)	0.000
SC1	0.488 (0.01)	0.000
SC2	0.084 (0.01)	0.000
UA1	0.393 (0.01)	0.000
UA2	0.197 (0.01)	0.000
PD1	0.563 (0.02)	0.000
PD2	0.309 (0.01)	0.000
AD1	0.538 (0.01)	0.000
AD2	0.205 (0.01)	0.000
Pseudo R^2	0.1736	
AIC	67271.21	
Log-likelihood	-33625.61	
MAE	0.058	
RMSE	0.0745	

The coefficients for omitted categories (level 3) can be calculated as the negative summation of non-omitted variables' coefficients. For example, β for MO3= - (0.618+0.351) = - 0.969.

In the interaction-effects model, all main effects were statistically significant at the 95% level (Table 3).

Table 3. Parameter estimates for interaction-effects model based on discrete choice (DC) data

	Interaction-effects estimates	
	β (SE)	P-value
MO1	0.636 (0.01)	0.000
MO2	0.329 (0.01)	0.000
SC1	0.489 (0.01)	0.000
SC2	0.077 (0.01)	0.000
UA1	0.397 (0.01)	0.000
UA2	0.187 (0.01)	0.000
PD1	0.572 (0.02)	0.000
PD2	0.291 (0.01)	0.000
AD1	0.550 (0.01)	0.000
AD2	0.194 (0.01)	0.000
MO× SC (Likelihood value)	98.30	0.000
MO1×SC1	0.104 (0.01)	0.000
MO1×SC2	0.000 (0.01)	0.989
MO2×SC1	-0.043 (0.01)	0.002
MO2×SC2	0.045 (0.01)	0.001
MO× UA (Likelihood value)	36.77	0.000
MO1×UA1	0.008 (0.01)	0.566
MO1×UA2	-0.003 (0.01)	0.831
MO2×UA1	0.043 (0.01)	0.001
MO2×UA2	0.028 (0.01)	0.019
MO× PD (Likelihood value)	36.81	0.000
MO1×PD1	0.083 (0.01)	0.000
MO1×PD2	-0.038 (0.01)	0.002
MO2×PD1	-0.048 (0.01)	0.001
MO2×PD2	0.032 (0.01)	0.022
MO× AD (Likelihood value)	11.17	0.025
MO1×AD1	0.001 (0.01)	0.958
MO1×AD2	-0.027 (0.01)	0.057
MO2×AD1	0.017 (0.01)	0.207
MO2×AD2	0.021 (0.01)	0.102
UA× SC (Likelihood value)	29.60	0.000
UA1×SC1	0.011 (0.01)	0.412
UA1×SC2	0.008 (0.01)	0.563
UA2×SC1	0.054 (0.01)	0.000
UA2×SC2	-0.018 (0.01)	0.172

SC× PD (Likelihood value)	24.74	0.000
SC1×PD1	0.064 (0.01)	0.000
SC1×PD2	-0.030 (0.01)	0.035
SC2×PD1	-0.021 (0.01)	0.122
SC2×PD2	0.027 (0.01)	0.036
SC× AD (Likelihood value)	29.89	0.000
SC1×AD1	0.053 (0.01)	0.000
SC1×AD2	-0.022 (0.01)	0.083
SC2×AD1	-0.032 (0.01)	0.023
SC2×AD2	0.056 (0.01)	0.000
UA× PD (Likelihood value)	65.43	0.000
UA1×PD1	0.040 (0.01)	0.003
UA1×PD2	-0.047 (0.01)	0.001
UA2×PD1	0.055 (0.01)	0.000
UA2×PD2	0.010 (0.01)	0.411
UA× AD (Likelihood value)	12.87	0.012
UA1×AD1	-0.010 (0.02)	0.536
UA1×AD2	-0.002 (0.01)	0.864
UA2×AD1	0.006 (0.01)	0.676
UA2×AD2	0.035 (0.01)	0.005
PD× AD (Likelihood value)	16.41	0.003
PD1×AD1	0.052 (0.01)	0.000
PD1×AD2	-0.023 (0.01)	0.084
PD2×AD1	-0.009 (0.01)	0.501
PD2×AD2	0.014 (0.01)	0.271
Pseudo R^2	0.1772	
AIC	67061.48	
Log-likelihood	-33480.74	
MAE	0.053	
RMSE	0.0673	

The coefficients for omitted categories (level 3) can be calculated as the negative summation of non-omitted variables' coefficients.
β for MO3= - (0.636+0.329) = - 0.965. β for interaction MO3×SC1= - (βMO1×SC1+ βMO2×SC1) = - (0.104-0.043) = -0.061.

Inclusion of all second-order interactions simultaneously resulted in a statistically significant improvement of model fit (log-likelihood ratio test: LR chi^2 (40) =289.74, P-value =0.00). Moreover, all ten pairwise interactions between attributes are significant. The interaction term consisting of mobility and self-care is the most salient one since its likelihood ratio test statistic is the highest (LR = 98.3) and the associated P-value is very low. The lowest likelihood ratio test statistic (LR = 11.17) was shown for the interaction of mobility with anxiety/depression (Table 3). However, inclusion of all second-order

interactions improves the fit only slightly based on the indicators of pseudo R^2 and AIC. The improvement of model fit by including interaction-effect based on pseudo R^2 was modest (rise from 0.174 to 0.177). Similar results were found for the AIC, whereby the lower AIC indicates better model fit (67271.2 for the main-effects model, 67061.5 for the interaction-effects model). The measures of model accuracy RMSE and MAE indicated in favor of the model with interactions in terms of predicting accuracy. Health states and predicted values from the main effect model and interaction effect model were plotted (Figure 2 and 3), and it was demonstrated that the interaction effects model shows lower values than the main effects model on the entire range of health states. The maximum difference between the values produced by a main-effects and interaction-effects model is 0.129, while the average difference is 0.076.

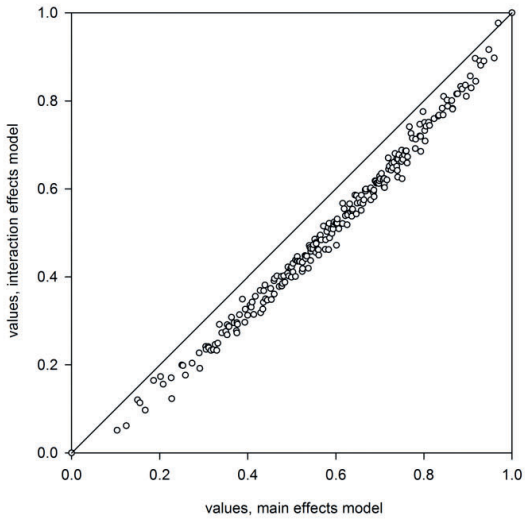

Fig. 2 Predicted values (scaled from 0 to 1) for 243 EQ-5D-3L health states based on the model with main effects and on the model including interactions

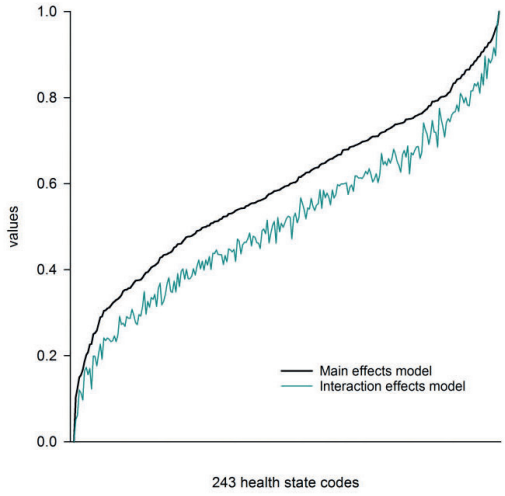

Fig. 3 Predicted values (scaled from 0 to 1) for 243 EQ-5D-3L health states based on the model with main effects and on the model including interactions, sorted by the values for the main-effects model

4. DISCUSSION

We have demonstrated the feasibility of deriving values for EQ-5D-3L states using a discrete choice model with all second-order interactions and efficient experimental design properties. It was shown that the effect of the health attributes is not simply additive. Interactions do contribute to the final estimated values for health states.

Most studies do not use all possible interaction terms but only those of interest [31]. For example, instead of including all second-order interactions, some studies [4, 12, 13] used one overall (omnibus) term (N3) to capture having severe (level 3) problems on at least one attribute. Other studies investigated the inclusion of a constant signifying any movement away from perfect health, or a D1 term (interaction term representing the number of movements away from perfect health due to having one or more attribute at level 2 or level 3) [32, 33]. These studies found little impact of interactions on the model fit. This is not surprising, as they were not designed to properly estimate all possible interactions between distinct health attributes. Intuitively, many combinations of health attributes are imaginable, in which case interactions would exist. For example, the ability to perform usual activities may depend on a person's mobility or feeling of pain/discomfort, since these attributes define and are integrated into usual activity.

The current study showed that although adding all possible second-order interactions improved the model fit, their inclusion improved the explained variance only slightly. The estimates were consistently lower moving downwards from level 1 (having no problems), which suggested declining values for health states associated with incremental

moves away from perfect health. The obtained estimates from interactions model are systematically lower than the estimates from the main effects model. Moreover, estimates were consistently negative, which suggested a declining marginal utility associated with additional shifts away from perfect health. The results of the present study demonstrated presence of interactions among the attributes in EQ-5D-3L, meaning that the effects of two or more health problems combined are stronger than the sum of the individual main effects. The same effect was investigated in the development of the HUI3 [7].

We found a number of quantitatively and statistically significant interactions among the attributes mobility, self-care, and pain/discomfort. The most salient one is between mobility and self-care; inclusion of this interaction term contributes more to model fit improvement than inclusion of the others. In the study of Mulhern et al. [34] investigating the interactions between the attributes of EQ-5D health state and duration, the interaction between pain/discomfort and duration showed the largest effect on values of health states, whereas the effect of interaction between mobility and duration was the lowest. In the study of Viney et al. [24], the weights for the attributes pain/discomfort, mobility, and self-care were larger. They also found that the following two interactions had the largest effects on the values of the health states: the interaction between mobility and self-care, the interaction between mobility and pain/discomfort. These findings concur with the current study. In the study of Jelsma & Maart [35] severe problems with mobility and pain/discomfort showed the largest significant effect on HRQoL as in the current study.

The present study has several strengths. An important one is the balance of design efficiency and response efficiency of our study. The design did not contain dominant pairs, and by implementation of two-level overlap response efficiency was reached. This made the response tasks easier, thereby reducing respondent fatigue [36-40]. Furthermore, a large sample was obtained, which made it possible to estimate and investigate all possible second-order interaction terms for the EQ-5D-3L. Many health states were included in the study, which increased the accuracy of the results and aided to estimate all possible second-order interactions.

The study has some limitations too. The first being, that no priors for interactions terms were used when constructing the experimental design. Priors were set to zero because none of the previous studies had investigated all possible second-order interactions for the EQ-5D-3L jointly. It may be argued that priors for interactions could have been achieved with a pilot study. However, this would have required redesigning 400 pairs of health states, terminating the sampling process, and rerunning the survey. Therefore, it was decided not to run a pilot, so the zero priors were set for interaction terms. A second limitation is that the results may be affected by the fact that the assessment of the EQ-5D-3L health states was performed by a sample of the general population. Newly developed ‘experience-based’ methods, which make use of patients who assess health-state descriptions and compare these to their own health condition [41], might reveal larger interaction effects. Another limitation is the absence of theoretical hypothesis for

testing specific interactions. However, the aim of our study was to investigate whether adding all second-order interactions in a model results in different estimates for the health states, and to test the feasibility of such a model, rather than testing specific interactions, such as the N3 term [14, 15, 42].

Testing specific interactions instead of all interactions could be beneficial to future research on the EQ-5D-5L, which has five instead of three levels for each of the five attributes, generating a much wider array of possible interactions. For this 5L version, testing all possible interactions could be troublesome due to the large number of parameters to be estimated and the very large sample size required. Therefore, theoretical knowledge and empirical evidence from the current study may be applied to select specific key interactions for further research. For example, the interactions among the attributes mobility and self-care, which appeared the most salient for the EQ-5D-3L could be investigated in the EQ-5D-5L.

To conclude, estimation of EQ-5D-3L states using statistical models comprising all second-order interactions is feasible. Health attributes are related to and dependent on each other, an assumption that has been confirmed by the significance of the interactions between the five attributes of the EQ-5D-3L. For the EQ-5D-3L, a value function based on interactions produces systematically lower values than a main-effects model. It seems that the simple main-effects model for the EQ-5D-3L instruments may not be sufficiently accurate to produce credible health-state values. However, the practical implications of the differences between values generated with or without interactions may be small, because differences between values for various health states seem more comparable.

5. REFERENCES

1. World Health Organization. The first ten years of the World Health Organization. Geneva: World Health Organization, 1958
2. Krabbe PFM. The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. San Diego, USA: Elsevier/Academic Press, 2016
3. Selivanova A, Krabbe PFM. Eye tracking to explore attendance in health-state descriptions. PLOS ONE 2018; 13(1): e0190111. https://doi.org/10.1371/journal.pone.0190111
4. Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete choice modeling for the quantification of health states: The case of the EQ-5D. Value Health 2010; 13: 1005-1013
5. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med 2001;33: 328-336
6. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21(2): 271-292
7. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002; 40(2): 113–128
8. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: How reliable is the relationship? Health Qual Life Outcomes 2009; 7: 27. doi: 10.1186/1477-7525-7-27
9. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 2014; 17(4): 445–453
10. Sullivan PW, Ghushchyan V. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. Med Decis Making 2006; 26:401-409
11. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires (2nd ed). New York: Oxford University Press, 1996
12. Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate societal health state utility values. J Health Econ 2012; 31:306–318
13. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ 2002; 11: 341–353.
14. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006; 15(10):1121-1132.
15. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997; 35(11): 1095-1108.
16. Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: Background, theories, assumptions and applications. Expert Rev. Pharmacoeconomics Outcomes Res 2013; 13(1): 93–108
17. Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care. 2008; 46(4): 357-365.
18. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law. 2009; 4: 527-546.
19. Thurstone LL. A Law of Comparative Judgment. Psychol Rev. 1927; 4: 273-286.

20. McFadden D. Conditional logit analysis of qualitative choice behavior.In: Zarembka P, ed. Frontiers in Econometrics. New York: Academic Press; 1974:105–142.
21. Krabbe PFM, Devlin NJ, Stolk EA, Shah KK, Oppe M, van Hout B, Quik EH, Pickard AS, Xie F. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014; 52(11): 935-943.
22. Kuhfeld WF. Marketing research methods in SAS: Experimental design, choice, conjoint, and graphical techniques. Technical Report, SAS Institute 2005 http://support.sas.com/techsup/technote/ts723.html
23. Box GE, Hunter JS, Hunter WG. Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition. Wiley, 2005.
24. Viney R, Norman R, Brazier J, et al. An Australian choice experiment to value EQ-5D health states. J Health Econ 2014; 23:729-742
25. Brazell JD, Louviere JJ. Length effects in conjoint choice experiments and surveys: An explanation based on cumulative cognitive burden. Department of Marketing, University of Sydney, 1998
26. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21: 271–292
27. Coast J, Flynn TN, Salisbury C, Louviere J, Peters TJ. Maximising responses to discrete choice experiments: A randomised trial. Appl Health Econ Health Policy 2006;5: 249–260
28. Hall J, Fiebig DG, King MT, Hossain I, Louviere JJ. What influences participation in genetic carrier testing? Results from a discrete choice experiment. J Health Econ 2006; 25: 520–537
29. Johnson R, Orme B. Getting the most from CBC. Sequim: Sawtooth Software Research Paper Series, Sawtooth Software, 2003.
30. De Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. The Patient. 2015; 8(5):373-384. doi:10.1007/s40271-015-0118-z.
31. Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J.International comparisons in valuing EQ-5D health states: A review and analysis. Value Health 2009; 12: 1194–1200
32. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: Development and testing of the D1 valuation model. Med Care 2005; 43: 203–220
33. Rand-Hendriksen K, Augestad LA, Dahl FA. A critical re-evaluation of the regression model specification in the US D1 EQ-5D value function. Popul Health Metr 2012; 10: 2. doi: 10.1186/1478-7954-10-2.
34. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using discrete choice experiments with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. Med Decis Making 2016; 37(3): 285-297
35. Jelsma J, Maart S. Should additional attributes be added to the EQ-5D health-related quality of life instrument for community-based studies? An analytical descriptive study. Popul Health Metr 2015; 13: 13. doi: 10.1186/s12963-015-0046-0
36. Johnson FR, Lancsar E, Marshall D. Constructing experimental designs for discrete-choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. Value Health 2013; 16: 3–13

37. Flynn TN, Bilger M, Malhotra C, Finkelstein EA. Are efficient designs used in discrete choice experiments too difficult for some respondents? A case study eliciting preferences for end-of-life care. Pharmacoecon 2016; 34: 273–284

38. Jonker MF, Attema AE, Donkers B, Stolk EA, Versteegh MM. Are health state valuations from the general public biased? A test of health state preference dependency using self-assessed health and an efficient discrete choice experiment. J Health Econ 2017; 26 (12): 1534-1547

39. Louviere JJ, Islam T, Wasi N, Street D, Burgess L. Designing discrete choice experiments: Do optimal designs come at a price? J. Consumer Res 2008; 35(2): 360–375

40. Maddala T, Phillips KA, Reed Johnson F. An experiment on simplifying conjoint analysis designs for measuring preferences. J Health Econ 2003; 12(12): 1035–1047

41. Krabbe, PFM. A generalized measurement model to quantify health: The multi-attribute preference response model. PLoS ONE 2013; 8(11): e79494. https://doi.org/10.1371/journal.pone.0079494.

42. Luo N, Johnson JA, Shaw JW, Coons SJ. A Comparison of EQ-5D index scores derived from the US and UK population-based scoring functions. Med Decis Making 2007; 27(3): 321-326

CHAPTER 3

Patients provide different values for health states than healthy respondents

Submitted to the European Journal of Health Economics

ABSTRACT

Background

Typically, the health-state values used in economic evaluations are derived from a representative community sample. Recently, investigators have turned to deriving values from patient preferences, reasoning that patients are better informed about or more able to imagine certain health states and to evaluate these.

Objectives

This study uncovers differences between people with disease experience and currently healthy respondents in the importance they assign to various EQ-5D-5L health domains and in the underlying health-state values.

Methods

Using efficient designs, 240 pairs of EQ-5D-5L health states were generated in a pairwise choice format. The online survey was administered to a large sample (3,068 respondents) consisting of healthy respondents and patients. Their responses were analyzed using a conditional logit model and the values assigned to each EQ-5D-5L health state were compared.

Results

The health-state values derived from healthy respondents were lower than those derived from patients. Compared to healthy individuals, patients attached more weight to self-care and less to mobility.

Conclusions

Differences in appraisal of health domains were found between individuals who had experienced disease and healthy individuals, resulting in dissimilarity in their health-state values. This study emphasizes the importance of values from respondents experienced certain health states or diseases.

1. INTRODUCTION

Conventionally, the values for health states used in economic evaluations are derived from a representative community sample, a practice based on several arguments [1, 2]. One reason to use community samples for valuing health states is that, as tax-payers and payers for health care (through insurance premiums), these persons are deemed to represent the general population [3, 4]. Another is the 'veil of ignorance' argument [2], whereby the general population is presumed to have never experienced the impaired health states under evaluation and to be blind to its own self-interest. Accordingly, representatives of the general population would embody principles of justice and equity, and, thereby, ensure a fair distribution of resources. Whereas both of the above reasons imply the use of healthy respondents, a general population sample is also expected to include patients. However, in reality it can consist mostly of healthy respondents, therefore, reflecting the preferences of healthy individuals mainly. In addition, the results derived from such a general population sample can dependent highly on the proportions of patients comprising this sample.

The arguments supporting the values elicited from the community sample related to the perspective of justice, the veil of ignorance, and different economic arguments of tax-payers have a normative character [2, 5, 6]. Moreover, these arguments were found to be not convincing enough as a rationale for strictly adhering to general public preferences [6]. Therefore, the question "Who should value health?" is still a subject of heated debate.

Alternatively, the rationale for using values based on assessments by patients is that patients are likely to be more adequately informed than healthy people or more adept at imagining certain health states and thereby be better positioned to make an informative judgment about the impact on perceived health of such states. That reasoning may be more compelling when respondents have to take into account severely impaired health states [7]. In short, people who have direct experience with impaired health may provide more reliable and valid health-state valuations [7]. Alert to concerns about validity, several researchers have suggested eliciting patient assessments rather than consulting a sample of unaffected members of the general population [3, 8, 9, 10]. However, it has also been argued that health states should not be valued by patients because they may have adapted their perspective and lowered expectations or a distorted understanding of full health [8, 11, 12]. Most of such arguments seem to be related largely to the specific valuation techniques (time trade-off, standard gamble) to derive health-states values. These techniques are more susceptible to various biases due to their procedures, particularly if the respondents have experienced impaired health conditions (i.e., are patients or former patients) [13, 14].

Discrepancies in values for health states between patients who have experienced a disease and the general public have been observed in several studies [12, 15, 16], fueling the debate on whose values of health to consider. However, comparison of the general population and patients does not always take into account the current experience with

certain diseases. Specifically, the proportion of healthy respondents and the proportion of patients experiencing a disease in the sample of the general population are not always accounted for. However, current experience with diseases should be accounted for and investigated, since it may influence the elicited values. For example, Dolan [17] demonstrated that experiencing poor health results in higher health values. Another example is the study of de Wit et al. [18], where patients assigned higher values to hypothetical health states than a healthy sample consisting of students did. However, that study applied conventional valuation methods such as standard gamble, time trade-off, or visual analogue scale, all of which are subject to bias and highly sensitive to framing effects [19, 20]. The study of Brazier et al. [21] suggested that the differences between the general population and patients can be explained by multiple factors such as difficulty of patients to imagine full health, poor descriptions of health states (for the general population), use of different internal standards, and adaptation or response shift. In the other study, the differences between patient and healthy population preferences were far smaller when methods other than the conventional valuation techniques were employed [10]. Therefore, the aim of this study is to detect whether people with experience of disease tend to assign different values to health states or different weights to certain health attributes than currently healthy respondents would do. According to Brazier et al. [21], the established literature compares values for hypothetical health states elicited from the general population against values of experienced health states elicited from patients, rather than against experienced health states elicited from the general population. However, nothing has been mentioned about the comparison of hypothetical health states elicited from the general population against hypothetical health states elicited from patients. Taken into account the importance of using a robust measurement framework as discussed in earlier studies [5, 10], the method of discrete choice modeling was chosen for the present study. To our knowledge, this is the first study to explore the differences between the patients' values for hypothetical health-states and those of currently healthy respondents using discrete choice modeling.

2. METHODS

2.1 Discrete choice model

Substantial differences between health values derived from the healthy population and patients respectively can be attributed to the way the values were measured. For instance, instead of asking patients to assess the full range of health states, from mild to severe, they were only presented with several specific disease-related health states in some studies [10, 22, 23]. Moreover, they were asked to assess health states one-by-one, without being offered a comparator health state. The methods used in these studies did not allow for discrimination (or trade-offs), even though discrimination by comparison of two or more entities is a crucial element in robust quantification of subjective phenomena

such as health values [24-26]. Therefore, the current study uses the method of discrete choice. Choice models are grounded in modern measurement theory and are consistent with the random utility model in economic theory [27]. Generally, the response task implies two or more health-state descriptions consisting of various attributes (domains, dimensions) of health, each with severity levels. Based on the descriptions, respondents express a preference for a health state they consider as better. Other authors have proposed and tested the validity and applicability of discrete choice models for health-state evaluations [28-32].

2.2 EQ-5D-5L

The EQ-5D was developed by the EuroQol Research Foundation (www.euroqol.org) as a relatively simple generic instrument that could be used in clinical studies to provide values of health states for use in economic evaluation [33]. The EQ-5D-5L descriptive system comprises five attributes: Mobility (MOB), Self-Care (SC), Usual Activities (UA), Pain/Discomfort (PD), and Anxiety/Depression (AD). Each attribute has five levels: no problems, slight problems, moderate problems, severe problems, and extreme problems [34]. EQ-5D-5L health states are defined by selecting one level from each attribute, with 11111 denoting perfect health (no problems with any attributes) and 55555 the worst possible health state (extreme problems with all attributes). Based on the responses to the instrument, a preference-based function can be applied to generate a single value for health.

2.3 Experimental design

The EQ-5D-5L contains five attributes with five levels each, yielding $5^5 = 3{,}125$ possible health states. Our experiment used the simplest discrete choice task: a paired comparison of two health states. Given the number of possible pairs (4,881,250), it is hardly possible to present all possible pairs to the respondents. Therefore, health-state pairs were carefully selected to generate a set that could provide sufficient information. Two issues were taken into consideration in selecting the health states: respondent fatigue and dominance within the pairs.

The credibility of responses can be compromised if respondents get bored or fatigued, a burden that reflects the complexity or number of tasks they are expected to carry out. Earlier studies suggest that up to 16 choice tasks would be acceptable to people and not affect their responses [35-37]. Accordingly, we offered each respondent a set of 16 choice tasks and reduced their complexity by implementing a two-level overlap in the health-state descriptions. Two-level overlap implies fixing two out of five attributes at the same level while varying the level of the other three attributes. Level overlap and other simplifying techniques were used in the previous study on similar issue [7]. In an instrument comprising overall five attributes with different levels, the design with varying levels of three attributes was considered optimal as this make the task much easier,

but at the same time still allows for a lot of variation in the attribute levels to maintain efficiency [38].

Dominance is a common problem in health-state valuation exercises since all attributes are ordered, and smaller health problems are always preferred to greater ones. Dominant pairs do not offer additional information but they do reduce design efficiency. Therefore, all combinations where one health state in a pair had all the attributes worse (or better) than the other health state in that pair were removed.

The number of non-dominant EQ-5D-5L health-state pairs with two-level overlap was set at 1,430,000. Out of this total, 240 pairs were selected and divided into 15 blocks. The sequence of selection and blocking was programmed using an efficient design (Ngene). An experimental design is considered efficient when the parameters are estimated with the lowest possible standard error. The creation of the design was based on an iterative procedure wherein designs are compared by a measure of statistical efficiency, namely D-error. The smaller D-error indicates a smaller standard error of the parameters under estimation. D-errors were computed on the basis of expected values of the model parameters. Efficient design in Ngene requires priors, which were derived from an earlier multinational study [32]. Instructions were given to all respondents at the beginning of the online survey (conducted using Sawtooth Software). In the current study the existing validated instrument EQ-5D-5L was used without additional bolt-ons and other modifications, therefore, no pretesting has been conducted.

2.4 Sample

According to the golden rule formulated by Johnson and Orme [39], $N>500c/(t\times a)$, where N (minimum sample size per a block of a survey), c (largest level), t (number of tasks), and a (number of alternatives). In the main-effects model for EQ-5D-5L in the present study, c=5, t=16, a=2, number of blocks is 15, and calculated minimum sample size is N>500*5/32*15 =1,171, although more sophisticated calculation procedure may be found [40]. In total, 3,442 respondents were invited to participate in a self-completion computer-based assessment to be conducted by an agency for market research (Survey Sampling International, Rotterdam, the Netherlands). The sample consisted of healthy respondents (individuals who currently do not experience disease) and patients (who do currently experience disease). We defined patients as individuals with actual diseases or serious complaints, therefore, we did not consider individuals with past experience of a disease. The rewards for participation were arranged via internal agreement between the participants and SSI. The patients sample was recruited by the marketing agency from a registered panel of patients. The self-reported current diagnosis of the recruited patients fell within the following categories: neck and back pain, abdominal pain, migraine, chronic pain, diabetes, heart disease, hearing or vision loss, asthma/COPD, eczema, mental health problems, stroke, rheumatism (osteoarthritis, arthritis), cancer, epilepsy, lung disease, and gastrointestinal disease. Besides these diagnoses, some patients had

sleep problems and fatigue. Each respondent was randomly assigned to one of the 15 blocks of the survey. No limits on time for completion were set.

2.5 Analysis

The data analysis was performed using a conditional logit model (asclogit, Stata 15.0) to estimate the regression coefficients of a value function for EQ-5D-5L, which included 20 dummy variables representing level 2, 3, 4, and 5 [32]. The regression coefficients were checked for logical ordering and significance in the discrete choice model for all attribute levels in EQ-5D-5L. Gender and four age groups (younger than 25, 26-35, 36-50, 51-65) were used as modifiers to see whether these background characteristics had significant effect.

Next, the values were calculated based on regression coefficients derived from the responses on discrete choice tasks from the two samples (healthy people vs. patients). The value for the health state is modeled as the product of the health-state characteristics (severity of an attribute, such as level 1 problems with mobility or level 2 problems with self-care) and health-state preference parameters (β). Based on derived coefficients, the comparison of relative attribute importance has been made for both samples. We used the original values derived with the choice model and rescaled them from 0.0 (worst health state) to 1.0 (best health state). This was done for graphical demonstration whether the health-state values in the healthy population differ from the health-state values in the patient's population within the whole range of possible states (Sigma Plot 13.0). It was decided not to pool the samples with subsequent creation of the group dummy variable, since this seems mathematically invalid (the two populations may have a different preference scale). Therefore, the estimation of interaction effects between the group dummy and the attributes was not performed.

The EQ-5D-5L values of healthy respondents currently not experiencing a disease were compared with the values of patients (i.e., individuals currently experiencing a disease) by means of a Kernel density graph (Stata 15.0) illustrating the distributions of values [41].

3. RESULTS

3.1 Sample

A sample of 3,442 respondents was registered to complete the survey. Responses from healthy respondents (i.e., people with no disease diagnosed) were collected from the Dutch sample in September - October 2016. Responses from patients were collected in January - February 2017. Once the data had been gathered, however, we had to exclude 289 respondents due to unreliable responses, in light of a completion time under two minutes for 16 choice tasks. In addition, we excluded 85 respondents who chose only the left-side alternatives (or only the right-side alternatives) for the whole survey. In the final step, 1,221

healthy respondents and 1,847 patients were included in the analysis (Table 1). The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Table 1. Respondents' characteristics

Characteristics	Healthy respondents completed EQ-5D-5L N=1,221	Patients completed EQ-5D-5L N=1,847
Male, N (%)	580 (48)	739 (40)
Age, mean(SD)	48.2 (14.3)	48.8 (14.1)
Age group, N (%)		
18-25	67 (12)	74 (10)
26-35	59 (10)	83 (11)
36-50	135 (23)	162 (22)
51-65	319 (55)	420 (57)
Female, N (%)	641 (52)	1,108 (60)
Age, mean(SD)	42.1 (14.4)	44.7(13.5)
Age group, N (%)		
18-25	121 (19)	135 (12)
26-35	115 (18)	171 (15)
36-50	179 (28)	363 (33)
51-65	226 (35)	439 (40)
Diseases, N (%)		
No diseases	1,221 (100)	-
Neck- and back pain	-	769 (42)
Fatigue	-	594 (32)
Sleep problems	-	458 (25)
Pain (abdomen, migraine, chronic, etc.)	-	391 (21)
Rheumatism (osteoarthritis, arthritis)	-	367 (20)
Asthma/COPD	-	332 (18)
Hearing or vision loss	-	309 (17)
Mental health problems	-	288 (16)
Eczema	-	281 (15)
Diabetes	-	256 (14)
Heart disease	-	177 (10)
Gastrointestinal disease	-	129 (7)
Lung disease	-	88 (5)
Cancer	-	83 (4)
Stroke	-	55 (3)
Epilepsy	-	36 (2)

Abbreviations: N-number, SD- standard deviation, COPD-Chronic obstructive pulmonary disease

3.2 Regression coefficients

No illogical ordering of level coefficients were found for healthy respondents (Table 2). For the patient sample, mobility was insignificant at level 2 (P-value= 0.746). Moreover, the coefficient of the attribute pain at level 2 had a positive sign, indicating the positive effect of having slight pain over having no pain. However, it was insignificant at the 99% confidence level (P-value=0.035). Similarly, level 2 of usual activities (P-value=0.032) was insignificant at 99% confidence level. Of greatest concern to patients as well as healthy respondents was having extreme problems with anxiety/depression. However, these samples differ in their appraisal of other attributes. For patients, the second most important concern was having extreme problems with self-care, while for healthy respondents extreme pain/discomfort was the second most important matter. Among patients, problems with mobility were considered the least important, whereas healthy respondents considered self-care the least important.

Table 2. Regression coefficients based on completed discrete choice EQ-5D-5L tasks for general population and patients

EQ-5D-5L	Healthy population, N=1,221	Patients, N=1,847
	β (SE)	β (SE)
Mobility		
No problems (reference)		
Slight problems	-0.362 (0.04)	-0.011 (0.03)***
Moderate problems	-0.542 (0.04)	-0.138 (0.03)
Severe problems	-1.261(0.05)	-0.768 (0.04)
Unable to	-1.585 (0.05)	-0.996 (0.03)
Self-care		
No problems (reference)		
Slight problems	-0.149 (0.05)	-0.196 (0.03)
Moderate problems	-0.317 (0.04)	-0.348 (0.03)
Severe problems	-0.999 (0.05)	-0.934 (0.04)
Unable to	-1.171 (0.05)	-1.141 (0.04)
Usual activities		
No problems (reference)		
Slight problems	-0.329 (0.05)	-0.071 (0.03)*
Moderate problems	-0.419 (0.04)	-0.157 (0.03)
Severe problems	-1.203 (0.04)	-0.716 (0.03)
Unable to	-1.602 (0.05)	-1.051 (0.04)
Pain/discomfort		
None (reference)		
Slight	-0.325 (0.05)	0.076 (0.04)*
Moderate	-0.580 (0.05)	-0.089 (0.03)
Severe	-1.615 (0.05)	-0.818 (0.04)
Extreme	-2.157 (0.05)	-1.136 (0.04)

Anxiety/depression		
None (reference)		
Slight	-0.363 (0.05)	-0.215 (0.03)
Moderate	-0.660 (0.05)	-0.402 (0.03)
Severe	-1.598 (0.05)	-1.076 (0.04)
Extreme	-2.291 (0.05)	-1.572 (0.04)
Gender (Male)	-0.044 (0.03)***	-0.030 (0.03)***
Age groups		
younger 25 (reference)		
26-35	-0.093 (0.05)*	-0.023 (0.04)***
36-50	-0.009 (0.04)***	-0.051 (0.03)*
51-65	-0.057 (0.03)**	-0.055 (0.02)*
Log likelihood	-10158.634	-17265.37
Wald chi2	4329.25	4882.41

All variables are significant at 99% confidence level, except for *P-value<0.05; **P-value<0.1; ***P-value>0.1 (insignificance)
Abbreviations: SE-standard error

3.3 Values

The healthy sample scored more health states as severe (range 0–0.5) and fewer health states as mild (range 0.7–1) (Figure 1). The predicted values of both samples were scaled (0.0 – 1.0) and plotted against each other (Figure 2). The values derived from the healthy respondents are lower than the values derived from patients across the whole range of EQ-5D-5L health states (Figure 2). The mean difference in values derived from healthy respondents and patients is 0.037 (sd = 0.034). The maximum difference between the healthy respondents' and patients' values is 0.115 for health state '21253'.

3.4 Age and gender modifiers

Gender differences were proved to be not significant, implying that health-state values did not differ between males and females for both samples. However, some differences between the age groups could be observed (Table 2). In the patients' sample respondents younger than 25 years and respondents over 50 years assigned lower values to health states resembling the response pattern of the healthy sample. However, for the sample of healthy respondents the differences between the age groups were not confirmed (not present).

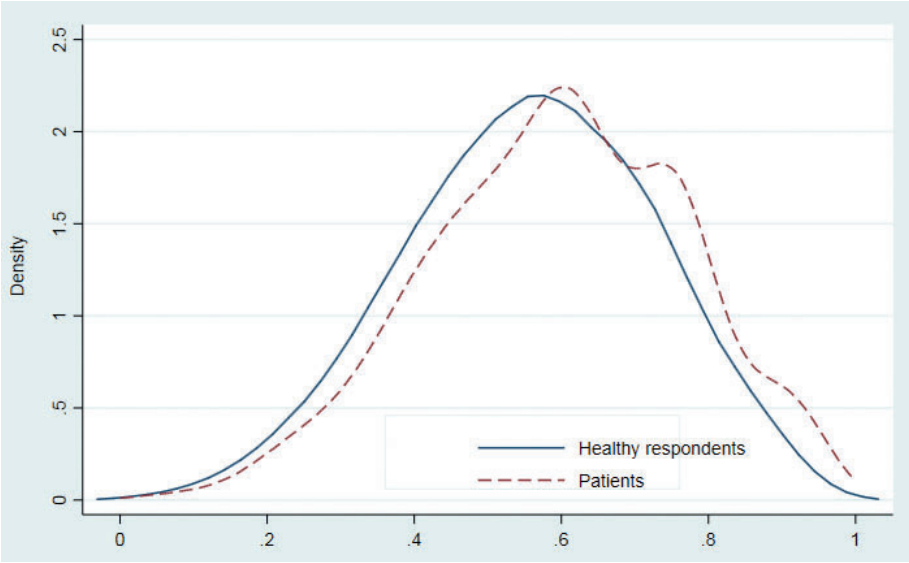

Fig. 1 Density plot of values for 3125 EQ-5D-5L health states elicited from healthy respondents and patients

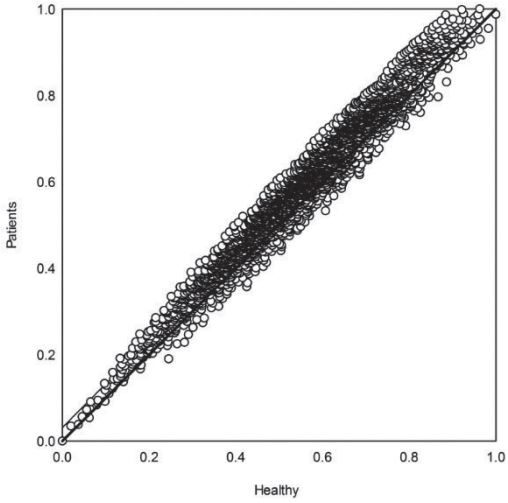

Fig. 2 Comparison of predicted values of 3125 EQ-5D-5L health states for healthy respondents versus patients

4. DISCUSSION

The present study assesses differences in health-state values between healthy individuals and patients. To avoid various confounding effects, we applied a straightforward and simple valuation method based on choice tasks. We found that the values conferred by healthy respondents are lower than those by patients. Our results suggest substantial differences in health-state evaluations between healthy and unhealthy individuals. In addition, we found that the values assigned to health states by older patients and younger patients are more similar to response patterns of healthy population, than the values of other patients' age groups.

The values assigned by respondents experiencing a disease are different than those of healthy respondents. A similar difference was observed by Dolan [17], who rejected the hypothesis that disease experience has no effect on health state evaluations. In fact, he demonstrated the converse: that experiencing poor health results in higher health values, which is confirmed by the current findings. Our study expands upon those findings by revealing differences between patients and the healthy population with respect to specific health attributes. For example, we found that patients assigned higher importance to extreme problems with self-care; it was the second most important attribute for patients but the least important one for healthy respondents. The study of Rand-Hendriksen et al. [4] found that self-care was the most important attribute for non-patients who had never experienced such health states under evaluation, but not for patients. In our study, however, the health states under evaluation can be hypothetical for patients as well, since there is no verification that the patients had experienced the health states they were evaluating.

Our study produced unexpected results for the patient sample, such as the insignificance of slight problems with mobility and the positive effect of slight pain/discomfort. Taking a closer look at the patients' responses, it is notable that slight problems with usual activities, mobility, and pain/discomfort were not significant at the 99% confidence interval. These attributes were the least important to the patients, most likely because patients tend to pay less attention to slight problems on attributes that are less important to them. As a result, they might consider these attributes as less meaningful in their decision-making process, i.e. dominated by problems in other domains.

The findings of the present study are not in line with those of Ogorevc et al., who used TTO valuation technique to explore the differences between preferences of the general public and patients towards EQ-5D-5L hypothetical health states [42]. The latter argued that patients considered problems with self-care as less important than the general population did. It should be noted that their study only investigated patients diagnosed with breast cancer (mostly women) or rheumatoid arthritis, which could influence the importance of attributes in the EQ-5D. For example, patients suffering from depression are more likely to assign greater importance to the attribute anxiety/depression. Therefore, by covering a wider range of diseases, our study would presumably reveal a

different pattern of attribute importance. Another explanation for the contrast between our results and those of Ogorevc et al. [42] lies in the different sampling approaches. Comparison of healthy respondents and patients, as in our study, can produce different results than comparison of the general population and patients. The problems with anxiety/depression were considered the most important by both samples in the present study. Similarly, the multinational study [32] and the Dutch study [41] demonstrated that most importance was assigned to anxiety/depression by the Dutch general population sample. The findings from the Swedish value set [43], based on the experienced health-state values using TTO and VAS elicited from the general population, showed the highest importance of anxiety/depression attribute, supporting the results of the present study). The study of Ogorevc et al. [42] demonstrated that problems with anxiety/depression are even more important for patients than for the general population.

The results of the study of Jonker et al. [7], using discrete choice modeling, are in line with the results of our study. Specifically, they revealed that respondents currently experiencing an impaired health level assign higher value to this level than respondents who currently experience a better health. However, the study of Jonker et al. [7] used a representative sample of the Dutch population without subdividing them into patients and non-patients. In addition, the study [7] accounted for self-rated health of the respondents, which is different from the current study. Similarly, the study of Brazier et al. [21] revealed that people who experienced an impaired health state tend to assign higher values on dysfunctional health states than members of the general population. The extent of this discrepancy tends to be much stronger when people value their own health state.

Conventionally, health-state values are derived from general population samples, which consist of both patients and non-patients and thus cannot be considered a 'perfectly healthy' sample [44]. In fact, the general population sample contains certain proportions of healthy individuals and patients. Therefore, the values elicited from the general population can be quite similar to the values of healthy respondents (in case the majority of the population is healthy) or to those of patients (in case the majority of the population has some experience of disease). It must be determined whether respondents from the general population are familiar with certain health states because they can be asked to value a state of health they have never experienced (although some might have). If neither the general population nor the patients have experienced that particular health state, the values they assign to it may be similar. In case the health states under evaluation are similar to the ones experienced by a respondent, greater differences may be expected.

Studies in Germany, the US, the UK, the Netherlands, and Sweden have put patient perspective into the spotlight [45, 46]. However, they apply divergent definitions of 'patients'. Versteegh and Brouwer [5] define them in health valuation perspective as individuals who are actually in the health state to be evaluated and thus have experience of this health state. The patients in our study are individuals with impaired health status

who were asked to place values on pairs of impaired health states that could be different from their own health state. We did not ask them to specify whether they had experience with the health states under evaluation. Therefore, the values were elicited from patients who might be not experienced with such health states. By the same logic, the healthy respondents could have been patients in the past and thus have experienced the health states under evaluation. It seems reasonable to assume that differences between the healthy population and patients are related to the experience with particular health states. Having health problems in general does not have a fixed impact on how hypothetical health states are valued, since different kinds of health problems are not universal. By contrast, the authors suggest that judgments of other (hypothetical) health states is largerly related to the specifics of the experienced health problems related to the presented hypothetical states. Therefore, the health states under evaluation might be not hypothetical for such respondents, which would advocate having the respondents value health states close to their own (current or past) health status. The approach of valuing own health states was supported by Brazier et al. [21] and Burström et al [43]. Such an approach would probably yield more informative valuations that most likely will show larger contrasts with values derived from healthy people. Analytically this could be employed with the specific preference-based novel measurement model based on item response theory [47].

In the current study the average impact across different perspectives on health, as determined by the specific disease experienced of the included patients. However, the averaging could potentially mask large differences between subgroups of patients and the general population. Therefore, one of the interesting questions for the future research could be investigation of differences in the subgroups of patients and the general population.

This study has several strengths. First, both of the study groups were drawn from large samples. Second, the same valuation method (discrete choice) and the same statistical analysis were used for both samples. This is an important advantage, since it was previously suggested that observable differences between the samples are attributable to the measurement methods used to qualify health states [10]. Third, an efficient design was used to maximize the precision of estimated regression coefficients, and respondent fatigue was avoided by applying level overlap.

There are some limitations to this study, too. Notably, it lacks respondents' actual self-reported health, scored according to the EQ-5D descriptive system. The data collection of the current study was designed as a part for a larger study, therefore, the self-reported health data could not be collected. Having information about the current (and past) health status of a respondent might provide insight into the interdependence between an experienced health state and the hypothetical ones [7]. Consequently, since there is no information about current (and past) health status of a respondent, generalizability

of the observed findings beyond the specific makeup of the sampled patient group may be limited.

By their nature, health-state values derived with choice models cannot be interpreted as absolute (cardinal) numbers due to two reasons. First, the best health state (full health) is dominant and cannot be used in the choice model as anchor. Second, the location of death is unknown since a 'death' option was not included. Consequently, DC models position health states on a scale between the best and the worst health states. Therefore, one of the main problems with choice models is normalizing its scale to a death-full health (0.0 – 1.0) scale. To solve this problem, a task extension or additional tasks should be included on the design, like death questions, duration on the health states or an accompanied TTO task. We did not normalize the values on death and full health; therefore, the values elicited in our study do not generate utilities and cannot be used to calculate QALYs. However, by applying the DC approach the generated values are far less affected by possible biases.

Differences were observed in the appraisal of health attributes between individuals who had experienced illness and those who had not. Patients attached more importance to self-care and less importance to mobility than healthy individuals. In conclusion, this study has detected clear differences in health-state values between healthy individuals and patients. Taking into account the revealed differences, we emphasize the importance of values from respondents experienced certain health states or diseases. Moreover, we recommend developing the practice of health-state evaluation from respondents assessing health states close to their own (current or past) health status, which can result in more realistic and informative values.

5. REFERENCES

1. Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the economic evaluation of health care programmes. Fourth ed. Oxford University Press; 2015.
2. Neumann PJ, Ganiats TG, Russell LB, et al. eds. Cost-Effectiveness in Health and Medicine. Oxford University Press; 2016.
3. Gandjour A. Theoretical foundation of patient v. population preferences in calculating QALYs. Med Decis Making 2010; 30 (4): 57-63.
4. Rand-Hendriksen K, Augestad L, Kristiansen IS, et al. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. Qual Life Res 2012; 21:1005–1012.
5. Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. Soc Sci &Med 2016; 165: 66-74.
6. PDUFA reauthorization performance goals and procedures fiscal years 2018 through 2022. Available at: https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm511438.pdf. Accessed April 5, 2018
7. Jonker MF, Attema AE, Donkers B, et al. Are health state valuations from the general public biased? A test of health state preference dependency using self-assessed health and an efficient discrete choice experiment. Health Econ 2016; 1-14.
8. Dolan P, Kahneman D: The interpretation of utility and their implications for the valuation of health. Econ J 2008; 118: 215-234.
9. De Wit GA, Busschbach JJV, De Charro F. Sensitivity and perspective in the valuation of health status: whose values count? Health Econ 2000; 9(2): 109–126.
10. Krabbe PFM, Tromp N, Ruers TJM, et al. Are patients' judgments of health status really different from the general population? Health Qual Life Outcomes 2011; 9:31.
11. Menzel P, Dolan P, Richardson J, et al. The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. Soc Sci & Med 2002; 55: 2149–2158.
12. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Qual Life Res 2003; 12(6): 599–607.
13. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ 2002; 11: 447–456.
14. Doctor JN, Bleichrodt H, Lin JH. Health utility bias: A systematic review and meta-analytic evaluation. Med Decis Making 2010; 30: 58-67.
15. Dolan P. Whose preferences count? Med Decis Making 1999; 19: 482–6.
16. Weyler E-J, Gandjour A. Empirical validation of patient versus population preferences in calculating QALYs. Health Serv Res 2011; 46: 1562–74.
17. Dolan P: The effect of experience of illness on the health state valuations. J Clin Epidemiol 1996; 49(5): 551-564.
18. De Wit GA, Busschbach JJV, De Charro F. Sensitivity and perspective in the valuation of health status: whose values count? Health Econ 2000; 9(2): 109–126.

19. Krabbe PF, Stalmeier PF, Lamers LM, et al. Testing the interval-level measurement property of multi-item visual analogue scales. Qual Life Res 2006; 15(10): 1651-61.
20. Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. Med Decis Making 1997; 17(2): 208-16.
21. John Brazier J, · Rowen D, Karimi M,· Peasgood T, Tsuchiya A, Ratcliffe J. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. Eur J Health Econ 2017; 19(6): 881-891.
22. McPherson K, Myers J, Taylor WJ, et al. Self-valuation and societal valuations of health state differ with disease severity in chronic and disabling conditions. Med Care 2004; 42:1143-1151.
23. Happich M, Von Lengerke T. Valuing the health state 'tinnitus': Differences between patients and the general public. Hear Res 2005; 207: 50-58.
24. Thurstone L.L. A Law of Comparative Judgment. Psychol Rev 1927; 4: 273-286.
25. McFadden D. Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (ed.), Frontiers in Econometrics. New York: Academic Press; 1974.
26. Krabbe PFM. The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. San Diego, USA: Elsevier/Academic Press; 2016.
27. Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: Background, theories, assumptions and applications. Expert Rev Pharmacoecon Outcomes Res 2013; 13(1): 93–108.
28. Hakim Z, Pathak DS. Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modeling. Health Econ 1999; 8(2): 103-116.
29. McKenzie L, Cairns J, Osman L. Symptom-based outcome measures for asthma: The use of discrete choice methods to assess patient preferences. Health Policy 2001; 57:193–204.
30. Ratcliffe J, Brazier J, Tsuchiya A, et al. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ 2009; 18: 1261–1276.
31. Bansback N, Brazier J, Tsuchiya A, et al. Using a discrete choice experiment to estimate societal health state utility values. Health Econ 2012; 31: 306–318.
32. Krabbe PFM, Devlin NJ, Stolk EA, et al. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care 2014; 52(11): 935-943.
33. Brooks R. EuroQol: the current state of play. Health Policy 1996; 37 (1), 53–72.
34. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011; 20 (10): 1727-1736.
35. Coast J, Flynn TN, Salisbury C, et al. Maximising responses to discrete choice experiments: A randomised trial. Appl Health Econ Health Policy 2006; 5: 249–260.
36. Hall J, Fiebig DG, King MT, et al. What influences participation in genetic carrier testing? Results from a discrete choice experiment. Health Econ 2006; 25: 520–537.
37. Viney R, Norman R, Brazier J, et al. An Australian choice experiment to value EQ-5D health states. Health Econ 2014; 23:729-742.

38. Maddala T, Phillips KA, Reed Johnson F. An experiment on simplifying conjoint analysis designs for measuring preferences. Health Econ 2003; 12(12): 1035–1047

39. Johnson R, Orme B. Getting the most from CBC. Sequim: Sawtooth Software Research Paper Series, Sawtooth Software, 2003.

40. De Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. The Patient. 2015; 8(5):373-384. doi:10.1007/s40271-015-0118-z.

41. Versteegh MM, Vermeulen KM, Evers SMAA, et al. Dutch tariff for the five-level version of EQ-5D. Value Health 2016; 19 (4): 343-352.

42. Ogorevc M, Murovec N, Fernandez NB, et al. Questioning the differences between general public vs. patient based preferences towards EQ-5D-5L defined hypothetical health states. Health Pol 2017. doi: 10.1016/j.healthpol.2017.03.011.

43. Burström K, Sun S, Gerdtham UG, Henriksson M, Johannesson M, Levin L-A, Zethraeus N. Swedish experience-based value sets for EQ-5D health states. Qual Life Res (2014) 23:431–442

44. Dolan P. The measurement of health-related quality of life. In: Culyer AJ, Newhouse JP. eds. Handbook of Health Economics. The Netherlands: North-Holland; 2000.

45. Devlin NJ, Appleby J. Getting the Most out of PROMs. Kings Fund, London; 2010.

46. Nilsson E, Orwelius L, Kristenson M. Patient-reported outcomes in the Swedish National Quality Registers. J Intern Med 2016; 279(2): 141-53.

47. Krabbe, PFM. A generalized measurement model to quantify health: The multi-attribute preference response model. PLoS One 2013; 8(11): e79494. https://doi.org/10.1371/journal.pone.0079494.

CHAPTER 4

Value judgment of new medical treatments: Societal and patient perspectives

Under Revision

ABSTRACT

Background

In many countries reimbursement for medical interventions is based on recommendations from advisory boards and committees that use multiple criteria in their assessment procedure. There is no agreement if the currently used criteria reflect the preferences of the general population, nor if theirs differ from patient preferences.

Objective

To determine the importance of certain criteria regarding new treatments, and explore whether there are differences in preference of these criteria between the general population and patients.

Methods

The study is based on the modern framework of discrete choice models where respondents are presented with judgmental tasks to elicit preferences. In this study, respondents were asked to choose between two hypothetical scenarios of patients receiving a new treatment. The scenarios graphically represent treatment outcomes and patient characteristics. Responses were collected from patients and the general population.

Results

Preferences were strongly and significantly affected by additional survival years, age at treatment, initial health condition, and the cause of disease. The analysis of the interaction terms showed significant differences between the importance that patients versus the general population assigned to the three criteria (‘age at treatment’, ‘initial health’, and ‘cause of disease’).

Conclusions

Overall, differences between patients and the general population are modest. However, apart from health gains, respondents thought the age of an individual, cause and burden of disease to be important factors in choosing which treatments should be provided to whom. This finding contrasts with many procedures used in the assessment of prioritizing new medical interventions.

1. INTRODUCTION

Many Western countries rely on the recommendations of advisory boards and committees for decisions on reimbursement for new drugs and other medical interventions. These independent parties assess the available evidence to determine whether the innovation offers added value to patients and society at large. In England, guidance on cost-effectiveness and clinical relevance is provided by the NICE (National Institute for Health and Care Excellence), which recommends quality-adjusted life years (QALYs) as the outcome measure in health benefit assessment [1]. In the USA the emphasis is placed on the patient-centered comparative effectiveness of existing medical interventions, providing specific criteria for health outcomes measures by which patient subpopulations can be accounted for and evaluated in different types of research [2]. The assessment procedure in one West-European country, the Netherlands, is expanded upon below [3-5].

In the Netherlands, the assessment of insured care is performed by the Appraisal Committee (Advies Commissie Pakket, ACP) of the National Health Care Institute (Zorginstituut Nederland, ZiN). The main criteria it uses to assess the therapeutic and societal value of drugs and other health interventions are necessity, efficacy, cost-effectiveness, and feasibility. Performance on these four criteria is assessed to decide whether the new intervention warrants incorporation in the national health insurance package. The four criteria are largely based on the scheme of the report by the Commission Dunning [6]. However, some sub-criteria can have a strong impact on the final decision to reimburse health interventions [7]. For example, despite low cost-effectiveness, reimbursement may be considered when no other medical intervention or treatment is available. Other arguments may overrule the application of a particular criterion: dealing with an orphan drug, posing a clear risk to other population groups (infections, anti-conception, addiction), or dealing with patients in severe condition (burden of disease).

The assessment procedure is not straightforward, as there is dependency and interconnectivity between several (sub) criteria. The Appraisal Committee also needs to take societal value judgments into account. These judgments are not solely based on clinical relevance and cost-effectiveness but also on equity and moral values. For instance, social justice and distributive justice are associated with a fair allocation and distribution of goods (e.g., medical treatments).

Given a diversity of possible influential criteria, it is difficult to design a general methodology to gather information for setting priorities. Many national and individual studies have examined the principles underpinning the assessment of health interventions [8-12]. Overall, results of these studies suggest heterogeneity of the identified criteria. For example, the study by van Exel [8] demonstrated the variety of existing viewpoints that could be observed in the society, and emphasized that no single decision rule can satisfy all equity principles and viewpoints simultaneously. The systematic review of Gu et al. [9] highlighted a large degree of variation in both methods and empirical results of the health care priority setting studies. For example,

the following criteria appeared in their systematic review and were covered by the majority of studies: age, severity, lifestyle/self-induced illness, size/ distribution of health gain, prevention versus cure, components of health gain, cost of treatment, end of life, and other contextual criteria and other characteristics of beneficiaries of treatments. Convergence among decision-makers on the relevance of criteria has been found by Tanios et al. [12], and some of the discrepancies found are strongly related to contextual factors. Thus, there is no consensus on the core criteria by which to value health interventions, and some authors even assert that the fundamental principles are poorly defined [13-15].

Another concern refers to the question 'whose values to use for priority setting' (general public or patients), which has been raised by the earlier studies [16-17]. For the majority of instruments assessing HRQoL, the values they produce are derived from a representative community sample [18]. Being tax payers, the general public are assumed not to serve own self-interest and, therefore, to embody principles of justice and equity. However, it is reasonable to assume that in many situations healthy subjects may be inadequately informed or lack sufficient imagination to make an appropriate judgment about the impact of hypothetical health states on their quality of life [19]. Many researchers claim that individuals are the best judges of their own HRQoL. They are likely to be more adequately informed than healthy people or more adept at imagining certain health states. Therefore, in the opinion of those researchers, it is the patients' judgments that should be elicited to obtain values for health states [20-23]. That reasoning may be more compelling when the respondents have to take into account severely impaired health states, since people who have direct experience with impaired health may provide more reliable and valid health-state valuations [24].

Better insight of the relative importance of core criteria might help to make policy assessment procedures more straightforward and transparent. Therefore, the aim of this study was to investigate the importance assigned to certain selected criteria from a societal and patient perspective. For this an experimental study was devised, in which the health contribution of new medical treatments was assessed.

2. METHODS

2.1 Selection of method and criteria

Discrete choice (DC) modeling is a widely used technique to elicit personal and societal preferences in health-valuation studies [25, 26]. The statistical literature classifies it within the modern framework of probabilistic discrete choice models that are consistent with economic theory (i.e., the random utility model) [27-32]. All DC models establish the relative merit of one phenomenon based on its relative attractiveness. This technique requires participants to make choices among two (paired) or more presented scenarios (choice tasks) described by the means of specific attributes with certain levels. However, careful

selection and identification of most important and informative attributes (and their levels) are needed to enable a respondent to process them without being fatigued. In the present study the attributes included in paired scenarios represent criteria, therefore, careful selection is essential to ensure that all, or at least the most prominent, aspects of the decision-making process are captured. Therefore, a rigorous literature review was conducted to extract a set of 25 criteria that reflect societal concerns about treatment effectiveness and equity considerations [33]. However, presenting all criteria in a single choice task would be too demanding for respondents, and many of the criteria are only applicable in very specific circumstances. Therefore, out of these criteria those were selected to construct hypothetical scenarios, which patients and the general population respondents could understand without additional information. The limited amount of selected criteria should reflect the most crucial characteristics of the treatments (such as health gains). Additionally, the criteria should reflect the characteristics of the potential recipients of the new treatment to assess its necessity depending on the burden and cause of the disease. Such scenarios reflect:

a) crucial outcomes of the new and standard treatments: change in health-related quality of life (HRQoL) after a new treatment, gain in life years after a new treatment, change in HRQoL after a standard treatment (if it exists and is accessible), gain in life years after a standard treatment (if it exists and is accessible)
b) crucial characteristics of the patient: age, initial HRQoL to reflect burden of disease, and cause of the acute event associated with either accident, genetics or unhealthy lifestyle

2.2 Scenarios

Each scenario covered a health condition before intervention, the effect of the available standard treatment, and the effect of a new treatment. Instead of conventional written in text descriptions, the scenarios consisted of graphical representations, which are presumably easier to comprehend and reduce the framing bias [34]. Pilot testing of the survey was performed, where respondents (n = 8, contacted face-to-face in the university of Groningen and online in April 2015) indicated that the task did not cause difficulties with understanding and admitted the visual attractiveness of the study design. Also some suggestions were given to improve the task instruction. The explanations of the initial health state, health gains after the new treatment and health gains after the standard treatments are presented in Figure 1.

2.2.1 Initial health state

The hypothetical patient's initial health state (the state prior to the acute event) was captured by age at onset and initial HRQoL (Figure 1a). Age at onset was categorized as 25, 50, or 75 years given the following assumptions: age 25 is considered a relatively healthy time of life, associated with optimal level of abilities, general intelligence and productivity [35]; age 50 is when illnesses or symptoms first arise and health graduately

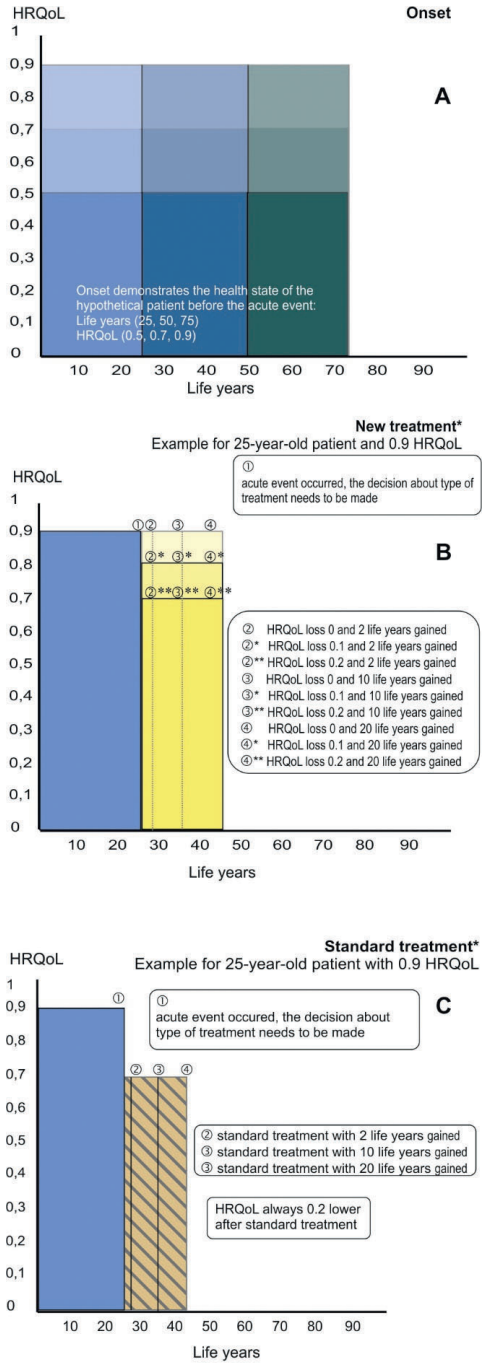

Fig. 1 Explanation of various options in the scenarios: (A) Possible health states before onset; (B) Example of new treatment for 25-year-old patient with 0.9 HRQoL ; (C) Standard treatment for 25-year-old patient with 0.9 HRQoL

starts to reduce [36]; age 75 presents the fastest growing section of populations in the developed world, and comprises those with substantial health problems [37]. Initial HRQoL was categorized as 0.5, 0.7, and 0.9 on a scale from 0.0 to 1.0, where 0.0 stands for the worst health and 1.0 for perfect health (explained in introduction of survey). The cut-off points were based on practical considerations. Perfect health is nearly impossible; in practice, 0.9 is as close to full health as one gets. Some health problems that limit normal functioning are common among patients whose HRQoL is 0.7, such as moderate angina [38]. Poor health is indicated by 0.5, which is seen among severely ill patients [38].

2.2.2 New and standard treatment

We assumed that HRQoL after any treatment could not exceed HRQoL before the acute event [39]. In other words, HRQoL after the acute event could not exceed the initial HRQoL, which is the prevailing case for most diseases and injuries. HRQoL is located within a range from 0.0 to 1.0, therefore, realistic changes (or decrements) in HRQoL had to be chosen. Thus, a change in HRQoL after the new treatment could be zero (thus, maintained at the previous level), decrease slightly (-0.1), or decrease substantially (-0.2). For the HRQoL change after standard treatment, we assumed a decrease of only 0.2. If the standard treatment resulted in higher HRQoL than the new one, a new treatment would not be needed. Thereby, we endorse the superiority of the new treatment over the standard one.

The gain in life years was categorized as 2, 10, and 20 additional years after undergoing a new treatment (Figure 1b). Standard treatment may not exist for some diagnoses, for instance, terminal cancer, Parkinson's disease, or Alzheimer's disease [40-42]. In other instances, standard treatment may exist but be inaccessible. It may be excluded from the state funding program or from the health insurance coverage, making it inaccessible for most patients. To account for the cases of standard treatment non-existence or inaccessibility, the gain in life years from standard treatment was set to zero. The gain was defined as 2, 10, and 20 years after receiving standard treatment, if existing and accessible (Figure 1c). In cases of inaccessible or non-existent standard treatment, the area of standard treatment gains did not appear in the scenario graph. The 2, 10, or 20 life years gained was assumed to represent possible life-extension effects. These gains were chosen to be evenly distributed across a plausible range of plausible combinations with maximum age, as Skedgel et al. [43] suggested. A minimum gain of 2 years was assumed to avoid comparisons with immediate death. Likewise, a median survival of 2 years is representative for a number of malignant diseases. Ten years of life gain was considered realistic for taking a number of chronic health states, such a heart failure in mind. A maximum gain of 20 years was assumed to yield realistic scenarios when combined with a maximum age of 75. These choices implied that a gain in life years after standard treatment could not exceed the gain after the new treatment.

2.3 Choice tasks

The respondents were given explanations of the paired scenarios and instructions on how to proceed with the task. Combinations of criteria were presented in three steps: first, the initial health state; second, the effects of the new treatment; third, a comparison between the new and the standard treatment (actual choice task). The patient's age and HRQoL before the acute event were shown on the x- and y-axis, respectively, while the benefits of the new and standard treatments were shown as color-coded and shaded areas in a plane. The term 'acute' was used to denote the sudden onset of a new disease, the sudden deterioration of an existing one, or the occurrence of an accident. Causation of the acute event was depicted by two icons that represented either unhealthy lifestyle elements, such as smoking/overweight, or external factors, such as genetic predisposition/accident. Based on presented information, the respondents had to decide which of the two scenarios for two hypothetical patients they preferred (Figure 2).

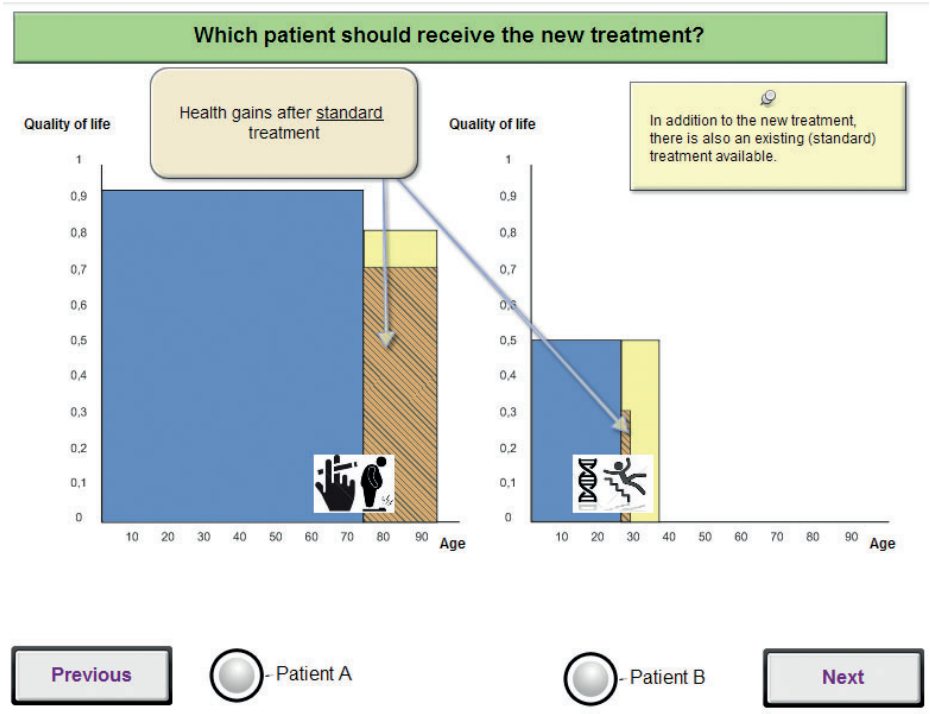

Fig. 2 Example of the discrete choice response task

2.4 Study design

In total, 200 pairs of scenarios were created. To diminish the burden on respondents and avoid fatigue, the set was divided into 20 blocks consisting of 10 choice tasks each. The tasks were selected from a whole range of possible combinations of paired comparisons

using efficient design (Ngene software, mnl model, null priors). An experimental design is called efficient if the parameters are estimated with the lowest possible standard errors. The design was based on an iterative procedure, where designs are compared by their D-error which is the measure of statistical efficiency [25]. The survey was programmed as a web-based experiment. The respondents were randomly assigned to one of the 20 blocks, meaning that each person completed 10 response tasks only.

2.5 Respondents

A representative sample on working age and gender drawn from the Dutch population and patient groups with a minimum age of 18 years was contacted by an agency for market research (Survey Sampling International, Rotterdam, The Netherlands). Individuals from patient groups were recruited and asked to self-report their diagnosis. The patients were asked to indicate whether they had any of the following types of diseases listed: diabetes, neck and back problems, heart disease, hearing or vision loss, asthma/COPD, eczema, mental health problems, stroke, rheumatism, cancer, epilepsy, lung disease, gastrointestinal disease, or other. Patients were allowed to report multiple diseases in case they had more than one diagnosed disease. For the general population, the sampling design did not verify whether potential respondents had been diagnosed with a disease. The Medical Ethics Review Committee at the University Medical Center of Groningen issued a waiver for this study, indicating that the pertinent Dutch Legislation (the Medical Research Involving Human Subjects Act) did not apply for this non-interventional study (METc 2014.181).

2.6 Analysis

Response data were analyzed in accordance with DC models by using the McFadden conditional logit model [44-45] with dummy-coded variables, representing the levels of the criteria (Stata, asclogit routine). Since the research question of the current study focuses on overall preferences and not on heterogeneity of preferences among respondents, the basic conditional logit model was considered as sufficient.

Deriving precise estimates for the (paired) scenarios requires a high number of respondents; according to Lancsar and Louviere [46], 20 respondents per questionnaire version is enough for the reliable model, but for significant post hoc analysis a bigger sample is required. In the present study, the minimal amount of respondents would have been 400 patients and 400 members of the general population, but having accounted for further analysis a larger sample was constructed.

The general population and the patients were compared in two ways. First, the importance assigned to the criteria was compared separately for both samples using the range method [47]. This method is used to compare the differences in preference weights between the best level of a criterion and the worst level of the same criterion. The calculated difference provides an estimate of the relative importance of that criterion

over the range of levels. Second, an analysis that included 2nd order interactions between respondent type and all criteria in the joint sample from analysis 2 was performed. Statistical significance of an interaction indicates that patients and the general population have different preferences for the criterion of interest.

3. RESULTS

3.1 Respondents

Data were collected from July 2015 till January 2016. In total 1,986 respondents from the general population and 2,256 patients were invited to participate. According to sociodemographic information provided by the SSI, both samples were representative of the Dutch population regarding sex and age. Some patients had been diagnosed with more than one problem, and the most common diagnoses were neck and back pain, diabetes, and asthma/COPD. Of the total amount of respondents registered for the survey by SSI (Table 1), the number of people who completed the survey and whose answers were included into analysis was 1,253 for the general population and 1,389 for the patients. The background characteristics have been collected only at the sample collection stage, and therefore, they are not presented for the analytical sample but only for the registered sample.

3.2 Importance of criteria

3.2.1 General population

For the general population, a gain in life years after the new treatment had the largest significant effect on the preferences (Table 2).

The second most influential criterion was age. They strongly favored treating a 25-year-old over a 50-year-old, and the effect was even stronger for a 75-year-old patient. When the cause of the acute event was related to an unhealthy lifestyle, a stronger negative effect on the preferences was found in comparison with a genetic or accidental cause. Finally, a higher initial HRQoL had a significant positive effect on the preferences. The capacity of the new treatment to maintain HRQoL at the same level as before the acute event had the weakest but still significant effect on their preferences. The characteristics of the standard treatment had no significant effect. The negative coefficient of 20 life years gained after standard treatment implied an unwillingness to give the new treatment to the hypothetical patient if standard treatment with 20 additional years of survival is available.

3.2.2 Patients

The patients showed a preference for a gain of life years after the new treatment, as the most important criterion. The cause of the disease was the second most important criterion, taking priority over age at onset (the overall weight of 0.23 for cause of disease exceeds the overall weight of 0.22 for age). Initial HRQoL and the ability of the treatment to

maintain the HRQoL at the same level as before the acute event were the least important criteria. The loss in HRQoL of 0.1 and the gain of 2, 10, and 20 life years with standard treatment were insignificant to the patients, relative to the effects of the new treatment.

3.3 Differences between the general population and patients

Overall, the importance of the criteria is almost identical for both samples when comparing the samples separately. The exceptions are the cause of the acute event and age: for the general population the importance of age is greater than the cause of the disease, while for patients the opposite holds. On the other hand, significant differences were found when analyzing preferences for specific criteria across the two samples. The latter analysis showed significant 2nd order interactions for the following criteria: age; initial HRQoL (0.9); cause of acute event; HRQoL change and life years gained after the new treatment (Table 3). The significance of the criteria age and cause of the disease

Table 1. Characteristics of the two study sub-samples

Characteristics	General population Overall registered N=1,986	Patients Overall registered N=2,256
Female, N (%)	1,104 (56)	1,239 (55)
Age, mean(SD)	46.6 (14.4)	47.8 (14.0)
Age group, N (%)		
18-24	286 (14)	244 (11)
25-34	194 (10)	223 (10)
35-44	240 (13)	281 (12)
45-54	485 (24)	580 (26)
Older 55	781 (39)	928 (41)
Diagnosed with*, N (%)		
Neck and back pain	-	995 (44)
Diabetes	-	736 (33)
Asthma/COPD	-	418 (19)
Mental health problems	-	383 (17)
Hearing or vision loss	-	370 (16)
Eczema	-	352 (16)
Rheumatism	-	335 (15)
Heart disease	-	302 (13)
Gastrointestinal disease	-	168 (7)
Cancer	-	153 (7)
Lung disease	-	86 (4)
Stroke	-	80 (4)
Epilepsy	-	53 (2)

*The total frequencies exceed 2,256 because some patients were diagnosed with more than one disease.

supported the results of the analysis based on the separated samples (Table 2). Finally, although the interactions between the sample type and the criteria HRQoL change and initial HRQoL 0.9 were significant, the size of these interactions did not change the order of the important criteria. Thus, an initial HRQoL in combination with a change in HRQoL after new treatment remained among the least important for both samples.

Table 2. Parameter estimates of the 6 criteria for the two sub-samples (based on completed surveys)

	General population (SE) N=1,253 Obs=47,756	Patients (SE) N=1,389 Obs=51,932
SCENARIO CRITERIA		
Patient characteristics		
Age		
Age 25 (reference)		
Age 50	-0.17(0.03)*	-0.10(0.03)*
Age 75	-0.67(0.04)*	-0.52(0.04)*
Initial Health-Related Quality of Life (HRQoL)		
HRQoL 0.5 (reference)		
HRQoL 0.7	0.28(0.03)*	0.24(0.03)*
HRQoL 0.9	0.42(0.03)*	0.28(0.03)*
Cause of acute event		
Accident, genetics (reference)		
Unhealthy lifestyle	-0.65(0.03)*	-0.55(0.03)*
New treatment outcomes		
*HRQoL change after new treatment (*ΔHRQoL)		
ΔHRQoL -0.2 (reference)		
ΔHRQoL -0.1	0.16(0.03)*	0.05(0.03)
ΔHRQoL 0	0.25(0.03)*	0.17(0.03)*
Life years gained after new treatment (LY_{new})		
LY_{new} 2(reference)		
LY_{new} 10	0.64(0.04)*	0.55(0.04)*
LY_{new} 20	0.94(0.04)*	0.84(0.04)*
Standard treatment outcomes* **		
Life years gained after standard treatment (LY_{st})		
Standard treatment unavailable (reference)		
$LY_{standard}$ 2	-0.04(0.03)	0.00(0.03)
$LY_{standard}$ 10	-0.04(0.03)	0.02(0.03)
$LY_{standard}$ 20	-0.10(0.05)**	0.01(0.04)
Goodness-of-fit	-14679	-16445
R-squared	0.1131	0.0863

*P<0.01, **P<0.05, *** -0.2 HRQoL is the fixed change after the standard treatment (7th criterion)

Table 3. Parameter estimates of the 6 criteria with interaction terms (criteria × sample type)

	General population and patients, model with interactions (SE), N=2,642, Obs=99,834
SCENARIO CRITERIA	
Patient characteristics	
Age	
Age 25 (reference)	
Age 50	-0.17 (0.02)*
Age 75	-0.67(0.02)*
Health-Related Quality of Life before onset (HRQoL)	
HRQoL 0.5 (reference)	
HRQoL 0.7	0.28 (0.03)*
HRQoL 0.9	0.42(0.02)*
Cause of acute event	
Accident, genetics (reference)	
Unhealthy lifestyle	-0.65(0.02)*
New treatment characteristics	
*HRQoL change after new treatment (*ΔHRQoL)	
ΔHRQoL -0.2 (reference)	
ΔHRQoL -0.1	0.16(0.03)*
ΔHRQoL 0	0.25(0.02)*
Life years gained after new treatment (LY_{new})	
LY_{new} 2(reference)	
LY_{new} 10	0.64 (0.03)*
LY_{new} 20	0.94 (0.03)*
Standard treatment characteristics* **	
Life years gained after standard treatment ($LY_{standard}$)	
Standard treatment unavailable (reference)	
$LY_{standard}$ 2	-0.04(0.02)
$LY_{standard}$ 10	-0.04(0.03)
$LY_{standard}$ 20	-0.10(0.04)*
RESPONDENT TYPE	
General population (reference)	
Patient	0.01(0.01)
INTERACTIONS BETWEEN THE SAMPLE AND THE CRITERIA	
Patient × Age 50	0.08(0.03)**
Patient × Age 75	0.15(0.03)*
Patient × HRQoL 0.7	-0.04(0.04)
Patient × HRQoL 0.9	-0.14(0.03)*
Patient × ΔHRQoL -0.1	-0.12(0.04)*

Patient × ΔHRQoL 0	-0.07(0.03)**
Patient × LY_{new} 10	-0.09(0.04)**
Patient × LY_{new} 20	-0.10(0.04)*
Patient × $LY_{standard}$ 2	0.04(0.03)
Patient × $LY_{standard}$ 10	0.05(0.03)
Patient × $LY_{standard}$ 20	0.12(0.05)
Patient × Unhealthy lifestyle	0.10(0.03)*
Wald chi2(25)	5657.42
Prob(chi2)	0.0000

*P<0.01, **P<0.05

*** *-0.2 HRQoL is the fixed change after the standard treatment (7th criterion)*

4. DISCUSSION

4.1 The aim of the study

The aim of the current study was to explore what importance the general population and patients assign to certain criteria, which reflected new and standard treatment outcomes and patient characteristics. The criteria were expected to capture societal concerns, particularly regarding treatment effectiveness and equity considerations. The study was designed to elicit possible differences in preferences between the general population and patients.

4.2 Literature review and discussion of results

The results of the analysis demonstrate the large importance of additional survival years due to the new treatment. Similar results were reported in a recent study of Skedgel [43] showing that respondents tend to favor scenarios where quality-adjusted life years (QALY) gain was the highest. However, it was also found in the present study that preferences regarding new treatments depend not only on the benefits of the treatments (gaining life years, maintaining HRQoL) but also on patient characteristics (such as the patient's age or initial health state). The earlier study in Germany examined the criteria for prioritizing health care based on patients' personal characteristics [48], found patients' severity of disease and HRQoL to be the most important attributes, while unhealthy lifestyle was found the least important. However, the authors [48] did not associate unhealthy lifestyle with the cause of disease as in the present study. The findings from the present study, as well as the study of Skedgel [43], suggest that preferences do not strictly follow only QALY-maximizing decision rules but incorporate both patient and treatment characteristics. The study of Baker [49] also found that age (younger population - children) and saving life even if HRQoL is low were important factors. However, strict maximization of health benefits was found to be important for a specific cluster of respondents in the study of Baker. Findings of the present study are partly in line with the findings of Baker [49] emphasizing the importance of younger age of a patient. Surprisingly, the availability and effect of a

standard treatment seems to have no effect on the appraisal of new treatments, which can be supported by the study results of Green et al. [26], who found the availability of 'other treatments' to be the least important attribute.

4.2.1 Fair innings argument

Our results can be placed in the context of 'fair innings' argument, expounded by Williams [50]. The argument is that everyone is entitled to a lifespan that is considered reasonable or 'fair'. The 'fair innings' argument takes the characteristics of the patients and the treatment (in terms of health gains) into account [51-52]. The argument covers the whole life span, whereby health in the past and actual age are accounted for, but also gain after treatment. According to Williams [50], gained life years for people having less than fair innings should be valued more highly than life years gained for people having fair innings or more. In our research this is affirmed by adding the high societal importance of age and initial health (i.e., health in the past and actual age) along with health gains after the new treatment (i.e., future gain). The study of Skedgel [43] from Canada and Sheffield (UK) [53] found strong preference towards younger individuals to receive a treatment. The study of Lancsar [54] in England found small preference weights attributed to the age at onset but larger weights attributed to the age at death.

4.2.2 Severity argument

The severity argument is widely used in the literature on social preferences [9, 16, 52, 55]. Shah [16] remarked that the most popular method to define severity is in terms of pre-treatment health state, which we incorporated in our study as initial HRQoL. But it needs to be noted that a lot of heterogeneity was observed in the definitions of severity and the study methods (personal trade-off, DC, social welfare function). Such heterogeneity may influence the outcomes of the studies. Nevertheless, Shah pointed out in his literature review that in the majority of studies respondents on the whole were willing to give priority to the severely ill. Skedgel et al. [53], as well as Shah [16], found that more severe patients were preferred over less severely ill. This is not in line with the findings of the current study, which revealed that the patients with higher initial HRQoL were favored over those with lower initial HRQoL. Similar results were found in the study of Wetering et al. [56] demonstrating higher preferences for treating persons who were already in a relatively good health state before treatment. In the later study the authors [56] used graphical representation of the scenarios, in which specific areas were depicted that indicated losses. This way of presenting could explain their findings. For example, some respondents just opted for the smallest health loss searching for the smallest area of losses in the graph [56]. In the present study the authors assumed the relative easiness and attractiveness of graphical format, improving it by visual aids, such as notes and balloons, popping up as additional help. However, it needs to be acknowledged that graphical representation might influence the respondents to opt for prioritizing scenarios

with the largest areas depicting initial health, taking into account age and initial HRQoL, and largest graph areas of gains in life years. In such a way the design might influence the decisions of those respondents who were focused on the sizes of graph areas, rather than on instructions and the implications of the scenarios.

4.2.3 Lifestyle-related cause of acute event

In addition, we show the importance of a lifestyle-related cause, a criterion rarely taken into consideration. The argument on individual responsibility for the cause of disease was raised in an earlier study of Singh et al. [57], which emphasized that the public gave higher priority to interventions for diseases where the patient has no control over the cause of the disease and lower priority to programs for illnesses that were "self-inflicted". In the other study investigating prioritizing health service innovation investments [58], the authors admitted that respondents did not prefer innovations targeting people with 'drug addiction' and 'obesity'. Although the respondents in our study tended to choose the alternatives with more additional survival years after the new treatment, they prioritized younger-aged patients with an accidental or genetic cause of the acute event and a higher initial HRQoL. The results confirm findings of other studies focused on societal preferences from various countries [43, 59-60]. For example, the findings of Luyten et al. [59] from Belgium suggested that higher preference was attributed towards individuals who did not cause their own illness. The general findings of Gu et al. [9] suggest the young are favored over the old, the more severely ill are favored over the less severely ill, and people with self-induced illness tend to receive lower priority. In those studies that considered health gain, larger gain is universally preferred, but at a diminishing rate.

4.2.4 Differences between the general population and patients

Additionally, our research focused on the differences between the general population and patients. The literature does not show consensus on the evaluation of differences between patients and the general population. For instance, the number of earlier studies noted dissimilarities in health-state values between these samples [60-63]. The current study, in contrast, detected no statistically significant difference between the general population and the patients, when taking all criteria into consideration. However, the analysis of specific interaction terms did reveal differences in the importance of criteria between the two types of sample. Specifically, the cause of the acute event was more important to patients than to the general population. The patients prioritized those who got the disease due to an accident or genetic predisposition. This demonstrates the importance that patients attach to taking responsibility for one's own health. By contrast, the general population assigned higher importance to age than to the cause of the acute event. The degree of importance was also enhanced by the significance of the interaction term between sample type and age. In addition, it needs to be mentioned that patients seem to have systematically lower parameter values, thus higher variance of error term,

which may likely reflect greater randomness of decisions compared to the representatives of the general population.

Although the two samples were representative of the Dutch population regarding sex and age, the respondents from the general population were on average slightly younger than the patients. Moreover, there were more people younger than 45 in the population sample than the patient sample. It is plausible that younger respondents in the general population sample would favor scenarios in which younger patients were presented. Luyten et al. [59] found that age-based preferences and accounting for lifestyle depends on the age and lifestyle of the respondents themselves. These findings could be a result of prioritizing personal interests, which is known as self-serving answers.

4.3 Limitations

This study has a few limitations, and it is useful to draw attention to them. First, it may be questioned whether the criteria incorporated in the study design adequately represent the ones people use to arrive at preferences for new health interventions. Since the comparison of options based on several criteria is a demanding task, we had to restrict the number of criteria we used to construct the scenarios. We included only the most relevant ones so as not to overload the respondent with information. Therefore, the authors are aware that the outcomes of the analysis could be influenced by the choice of criteria and criteria levels ranges. Choice of levels for attributes influences the range of estimated coefficients, which, consequently, influences the estimated overall weights of the attributes. Second, some researchers may argue that using graphs and icons to describe hypothetical states can cause framing bias. For example, the scale and format of the graphs and the design of the icons might affect the decision-making process of the respondents. We added written in text explanations and notes to support the graphical presentation, aids that reduce framing bias [34]. The earlier study in the area of discrete choice analysis suggested that graphical representation helps the respondent to understand the task better, while written in text explanations facilitate judgment [64]. Third, for the general population, the sampling design did not verify whether potential respondents had been diagnosed with a disease. Although we acknowledge that partially the general population consists of patients, the study did not aim to investigate the general population's proportions of healthy and non-healthy representatives.

4.4 Conclusion

To conclude, we found that five out of the six presented criteria affected preferences for specific treatment scenarios. These influential criteria are initial health, lifestyle-related characteristics, age of patient receiving the treatment, gain in life years, and change in HRQoL after the new treatment. However, patients and the general population were found to differ slightly in their ranking of age and cause of the acute event. The patient's age and initial health seem to be important factors when judging the value of new medical

treatments. Surprisingly, the cause of an acute event that calls for medical treatment seems to play a significant role in value judgments made by patients but a lesser one in those made by the general population.

There are no large differences in attributing value to specific combinations of health scenarios and treatment outcomes between patients and respondents from the general population. However, apart from health gains, respondents thought the age of an individual, cause and burden of disease to be important factors in choosing which treatments should be provided to whom. This finding contrasts with many procedures used in several countries in their assessment of prioritizing new medical interventions.

5. REFERENCES

1. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. Pharmacoecon. 2008; 26(9): 733 – 44.
2. Brixner DI, Watkins JB. Can CER be an effective tool for change in the development and assessment of new drugs and technologies? J Manag Care Pharm. 2012; 18(5): S6-S11.
3. Pakketbeheer in de praktijk. CVZ, Diemen; 2006.
4. Rechtvaardige en duurzame zorg. Advies uitgebracht door de Raad voor de Volksgezondheid en Zorg aan de minister van Volksgezondheid, Welzijn en Sport Zoetermeer; 2007.
5. Zinnige en Duurzame zorg. Advies uitgebracht door de Raad voor de Volksgezondheid en Zorg aan de minister van Volksgezondheid, Welzijn en Sport, Zoetermeer; 2006.
6. Commissie Keuzen in de zorg (Commissie Dunning), Kiezen en delen. Den Haag, SDU; 1991.
7. Brouwer WBF, Culyer AJ, van Exel NJA, Rutten FFH. Welfarism vs. extra-welfarism. J Health Econ. 2008; 27: 325-338.
8. Van Exel J, Baker R, Mason H, Donaldson C, Brouwer W, EuroVaQ Team. Public views on principles for health care priority setting: Findings of a European cross-country study using Q methodology. Soc Sci & Med. 2015; 126: 128-137.
9. Gu Y, Lancsar E, Ghijben P, Butler JRG, Donaldson C. Attributes and weights in health care priority setting: A systematic review of what counts and to what extent. Soc Sci & Med. 2015;146: 41-52
10. Kaplan RM. Value judgement in the Oregon Medicaid Experiment. Med Care.1994; 32(10): 975-88
11. Fischer KE. A systematic review of coverage decision-making on health technologies-evidence from the real world. Health Policy. 2012; 107(2-3): 218-230.
12. Tanios N, Wagner M, Tony M, Baltussen R, van Til J, Rindress D, Kind P, Goetghebeur MM; International Task Force on Decision Criteria.. International task force on decision criteria. Which criteria are considered in healthcare decisions? Insights from an international survey of policy and clinical decision makers. Int J Technol Assess Health Care. 2013; 29(4): 456-65.
13. Holm S. The second phase of priority setting. Goodbye to the simple solutions: the second phase of priority setting in health care. BMJ. 1998; 317(7164): 1000-2.
14. Culyer A. Need - is a consensus possible? J. Med. Ethics. 1998; 24(2): 77-80.
15. Cookson R, McCabe C, Tsuchiya A. Public healthcare resource allocation and the Rule of Rescue. J Med Ethics. 2008; 34(7): 540-4.
16. Shah KK. Severity of illness and priority setting in healthcare: a review of the literature. Health Policy. 2009; 93(2-3): 77-84.
17. Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. Soc Sci & Med. 2016; 165: 66-74.
18. Drummond MF, Sculpher MJ, Claxton K, Greg L. Stoddart, Bernie J. O'Brien, George W. Torrance. Methods for the economic evaluation of health care programmes. Fourth ed. Oxford University Press; 2015.
19. Krabbe PFM. The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. San Diego: Elsevier/Academic Press; 2016.

20. Gandjour A. Theoretical foundation of patient v. population preferences in calculating QALYs. Med Decis Making. 2010; 30 (4): 57-63.
21. Dolan P, Kahneman D: The interpretation of utility and their implications for the valuation of health. Econ J. 2008; 118: 215-234.
22. De Wit GA, Busschbach JJV, De Charro F. Sensitivity and perspective in the valuation of health status: whose values count? Health Econ. 2000; 9(2): 109–126.
23. Krabbe PFM, Tromp N, Ruers TJM, et al. Are patients' judgments of health status really different from the general population? Health Qual Life Outcomes. 2011; 9:31.
24. Jonker MF, Attema AE, Donkers B, et al. Are health state valuations from the general public biased? A test of health state preference dependency using self-assessed health and an efficient discrete choice experiment. Health Econ. 2016; 1-14.
25. Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. Value Health. 2010; 13:1005-1013
26. C. Green, K. Gerard. Exploring the social value of health-care interventions: a stated preference discrete choice experiment. Health Econ. 2009, 18(8): 951–976.
27. Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: Background, theories, assumptions and applications. Expert Rev Pharmacoecon Outcomes Res. 2013; 13(1): 93–108.
28. Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care. 2008; 46(4): 357-365.
29. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law. 2009; 4: 527-546.
30. Louviere LL, Woodworth G. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. J Mark Res.1983; 20(4): 350-367.
31. Marley AAJ, Louviere JJ. Some probabilistic models of best, worst, and best-worst choices. J Math Psychol. 2005; 49: 464-480.
32. Thurstone LL. A Law of Comparative Judgment. Psychol Rev. 1927; 4: 273-286.
33. Vermeulen KM, Krabbe PFM. Value judgment of health interventions from different perspectives: arguments and criteria. Cost Eff Resour Alloc. 2018; 16:16. https://doi.org/10.1186/s12962-018-0099-6.
34. Gamliel E, Kreiner H. Is a picture worth a thousand words? The interaction of visual display and attribute representation in attenuating framing bias. Judgm Decis Mak. 2013; 8(4): 482–491.
35. Skirbekk V. Age and Individual Productivity: A Literature Survey. Vienna Yearbook of Population Research. 2004; 2: 133-153.
36. Smith JP. Unraveling the SES: Health Connection. Population and Development Review. 30, Supplement: Aging, Health, and Public Policy. 2004: 108-132.
37. Matthews RJ, Smith LK, Hancock RM, Jagger C, Spiers NA. Socioeconomic factors associated with the onset of disability in older age: a longitudinal study of people aged 75 years and over. Soc Sci & Med. 2005; 61(7): 1567-1575.
38. Torrance GW. Utility approach to measuring health-related quality of life. J Chronic Dis. 1987; 40(6): 593-600.
39. Megari K: Quality of Life in Chronic Disease Patients. Health Psychol Res. 2013; 1(3): 27.

40. Lee YJ, Yang JH, Lee JW, Yoon J, Nah JR, Choi W-S, Kim C-m. Association between the duration of palliative care service and survival in terminal cancer patients. Support. Care Cancer. 2015; 23(4): 1057–1062.
41. Connolly BS, Lang AE. Pharmacological Treatment of Parkinson Disease: A Review. JAMA. 2014; 311 (16): 1670–1683.
42. Kumar A, Singh A, Ekavali. A review on Alzheimer's disease pathophysiology and its management: an update. Pharmacol Rep. 2015; 67(2): 195-203.
43. Skedgel C, Wailoo A, Akehurst R, Hon. Societal Preferences for Distributive Justice in the Allocation of Health Care Resources: A Latent Class Discrete Choice Experiment. Med Decis Making. 2015; 94-105.
44. McFadden D. Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (ed.), Frontiers in Econometrics. New York: Academic Press; 1974.
45. McFadden, D.: Economic choices. Am Econ Rev. 2001; 91(3): 351-378.
46. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics. 2008; 26 (8): 661–77.
47. Orme B. Interpreting the Results of Conjoint Analysis. In: Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research. Second Edition, Madison, Wis.: Research Publishers LLC; 2010.
48. Diederich A, Swait J, Wirsik N. Citizen Participation in Patient Prioritization Policy Decisions: An Empirical and Experimental Study on Patients' Characteristics. PLoS ONE. 2012; 7(5): e36824.
49. Baker R, Wildman J, Mason H, Donaldson C. Q-ing for health – a new approach to eliciting the public's views on healthcare resource allocation. J Health Econ. 2014; 23: 283–297.
50. Williams A. Intergenerational equity: an exploration of the 'fair innings' argument. J Health Econ. 1997; 6: 117-132.
51. Stolk EA, Pickee S, Ament AHJA, Van Bussbach JJ.Equity in health care prioritization: An empirical inquiry into social value. Health Policy. 2005; 74: 343-355.
52. Nord E. Concerns for the worse off: fair innings versus severity. Soc Sci & Med. 2005; 60: 257-263.
53. Skedgel CD, Wailoo AJ, Akehurst RL. Choosing vs. allocating: discrete choice experiments and constant-sum paired comparisons for the elicitation of societal preferences. Health Expect. 2015; 18(5): 1227–1240.
54. Lancsar E, Wildman J, Donaldson C, Ryan M, Baker R. Deriving distributional weights for QALYs through discrete choice experiments. J Health Econ. 2011; 30(2): 466–78.
55. Whitty JA, Lancsar E, Rixon K, Golenko X, Ratcliffe J. A systematic review of stated preference studies reporting public preferences for healthcare priority setting. Patient. 2014; 7(4): 365–86.
56. van de Wetering L, van Exel J, Bobinac A, Brouwer WBF. Valuing QALYs in Relation to Equity Considerations Using a Discrete Choice Experiment. Pharmacoecon. 2015; 33(12): 1289–1300.
57. Singh J, Lord J, Longworth L, Orr S, McGarry T, Sheldon R, Buxton M. Does Responsibility affect the public's valuation of health care interventions? A Relative Valuation Approach to Health Care Safety. Value Health. 2012; 15: 690–698.
58. Erdem S& Thompson C. Prioritising health service innovation investments using public preferences: a discrete choice experiment. BMC Health Serv Res. 2014; 14(1).

59. Luyten J, Kessels R, Goos P, Beutels P. Public Preferences for Prioritizing Preventive and Curative Health Care Interventions: A Discrete Choice Experiment. Value Health. 2015; 18: 224-233.

60. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Qual Life Res. 2003; 12: 599-607.

61. Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. J Health Econ. 2009; 18(3): 363–372.

62. Little MHR, Reitmeir P, Peters A, Leidl R. The Impact of Differences between Patient and General Population EQ-5D-3L Values on the Mean Tariff Scores of Different Patient Groups. Value Health. 2009; 17: 364 – 371.

63. Rowen D, Mulhern B, Banerjee S, Tait R, Watchurst C, Smith SC, Young TA, Knapp M, Brazier JE. Comparison of General Population, Patient, and Carer Utility Values for Dementia Health States. Med Decis Making. 2015; 35(1): 68–80.

64. Vriens M, Loosschilder GH, Rosbergen E, Wittink DR. Verbal versus Realistic Pictorial Representations in Conjoint Analysis with Design Attributes. J. Prod. Innovat. Manag. 1998; 15: 455-467.

CHAPTER 5

Eye tracking to explore attendance in health-state descriptions

Selivanova A, Krabbe PFM. Eye tracking to explore attendance in health-state descriptions. PLOS ONE 2018; 13(1): e0190111. https://doi.org/10.1371/journal.pone.0190111

ABSTRACT

Introduction

A crucial assumption in health valuation methods is that respondents pay equal attention to all information components presented in the response task. So far, there is no solid evidence that respondents are fulfilling this condition. The aim of our study is to explore the attendance to various information cues presented in the discrete choice (DC) response tasks.

Methods

Eye tracking was used to study the eye movements and fixations on specific information areas. This was done for seven DC response tasks comprising health-state descriptions. A sample of 10 respondents participated in the study. Videos of their eye movements were recorded and are presented graphically. Frequencies were computed for length of fixation and number of fixations, so differences in attendance were demonstrated for particular attributes in the tasks.

Results

All respondents completed the survey. Respondents were fixating on the left-sided health-state descriptions slightly longer than on the right-sided. Fatigue was not observed, as the time spent did not decrease in the final response tasks. The time spent on the tasks depended on the difficulty of the task and the amount of information presented.

Discussion and conclusion

Eye tracking proved to be a feasible method to study the process of paying attention and fixating on health-state descriptions in the DC response tasks. Eye tracking facilitates the investigation of whether respondents fully read the information in health descriptions or whether they ignore particular elements.

1. INTRODUCTION

Several preference-based measurement frameworks have been developed to quantify health conditions and express their quality [1, 2]. In general, preference denotes the relative 'desirability' of an object, and measures produced by preference-based methods are referred to as values or utilities. The health descriptions judged in a preference-based measurement framework usually comprise a set of distinct health attributes, each with a few levels to express their severity. The application of preference-based methods implies weighting these levels in multi-attribute health classification systems, such as the EQ-5D [3].

The core of a preference-based measurement framework consists of a response task comparing at least two objects. In the case of health these can be a pair of health descriptions (or combinations with duration of life or risk of immediate death). The respondents do not score the health attributes one by one but consider the whole set of attributes (i.e., health description), which requires reading and mentally processing all of the attributes presented. Subsequently, the complete description should be compared with another description or other information elements (e.g., time duration, risk of immediate death) in another description. After given descriptions are compared, a choice has to be made in favor of the most preferable one.

Conventionally, there were two preference-based methods commonly used for health state valuations: TTO (time trade-off) and SG (standard gamble). These methods – often called valuation techniques – proved to have substantial biases [4-8]. Moreover, the length and complexity of the tasks in these methods make it difficult for the respondents to perform them. For these and other reasons, attention is turning to new preference-based methods, in particular to the probabilistic discrete choice (DC) model introduced by McFadden [9]. The statistical literature classifies it within the modern framework of probabilistic discrete choice models that are consistent with economic theory (i.e., the random utility model). The underlying DC principle is that people's choices are based on the attractiveness of attributes [10, 11]. This method requires participants to make choices among two or more presented scenarios (choice tasks) described by the means of specific attributes with certain levels.

The choice processes can be described by the means of process tracing research [12]. Process tracing methods often examine the information individuals seek before making a choice and how that information produces a choice. The results of such tracing DC studies showed that respondents attend more attractive alternatives (no health attributes in this case) and important attributes, and this focus increases with practice [13]. Results also demonstrated that respondents making repeated DC tasks focus on the information that is most relevant to make a decision. In addition, the study of Orquin & Mueller Loose [14] rigorously evaluated several theories regarding the attention and eye movements during completing the choice tasks and confirmed the assumption that the favored alternative or most important attribute receives attention. Interestingly, the assumption of complete

information acquisition commonly assumed in preference-based methods in health-state evaluation (including DC), was rejected. Such an assumption implies that no information is disregarded and respondents pay attention to all information elements of the response task: the instructions, the full description of the health states, and other elements such as visual cues. However, this assumption was not directly verified in the health-state evaluation settings. Furthermore, the findings of aforementioned studies are not considered enough to verify the assumption, because they were not applied to the area of health-state evaluations, where the content is different and attributes are more interrelated.

Therefore, the current exploratory study investigates the process of paying attention to various information elements of a DC task in a health setting, such as: health-state descriptions presented as alternatives on the left or right side of the screen, or specific attributes describing the health-state. For this, eye tracking - a common research tool in marketing, cognitive science, human computer interaction, and psychology - was used to study the process of paying attention.

2. METHODS

2.1 Respondents

The eye-tracking study was performed in the Netherlands with 10 respondents who were either members of the general public, or PhD/Master students of the University Medical Center Groningen in contiguous scientific fields (Medicine, Medical Biology, Medical Microbiology). All respondents live in the province of Groningen, the Netherlands. A sample as small as five is often considered sufficient for qualitative and explorative studies [15, 16]. The respondents from the general public and students were personally contacted by the authors and invited to participate in the eye-tracking experiment. After receiving the verbal consent of participation, the time and place for the experiment were settled, and all the device installations and adjustments were fulfilled by the authors. The positioning between the respondent and the eye tracking device (Gazepoint GP3 binocular video-based system, 60Hz, 0.5-1 degrees precision) was settled as 65 cm, and the distance of 40 cm below the eye level according to the Gazepoint device setting instructions. In case the positioning is arranged differently, the precision of the tracker can diminish. The respondents were asked not to move their heads with large amplitudes to avoid the imprecise capture of the eye focus. The authors were within reachable distance to help with the exercises. The Medical Ethics Review Committee at the University Medical Center of Groningen issues waivers for this type of studies, indicating that the pertinent Dutch Legislation (the Medical Research Involving Human Subjects Act) do not apply to attitude and opinion studies, therefore, no Ethics Approval is needed for such studies.

2.2 Tasks

The eye-tracking experiment started with an eye-calibration procedure (9-point calibration on the black screen). Calibration is necessary to establish that the eye-tracking device captures the eye fixation precisely and to minimize deviations between the real focal point and the "tracked" one. After calibration, the respondents were allowed to begin the response tasks. In case that after the first calibration the eye-tracker was unable to detect the eye fixation precisely, the recalibration was performed until precision was reached.

Choice-based response tasks were presented as PowerPoint slides (font size 12) in the Dutch language. The layout was identical to the EQ-VT (EuroQol Valuation Technology) system and consisted of time trade-off (TTO, not part of this study) and DC tasks [17]. The DC response task called for a choice between two health states based on the verbal description of these states (Fig 1a).

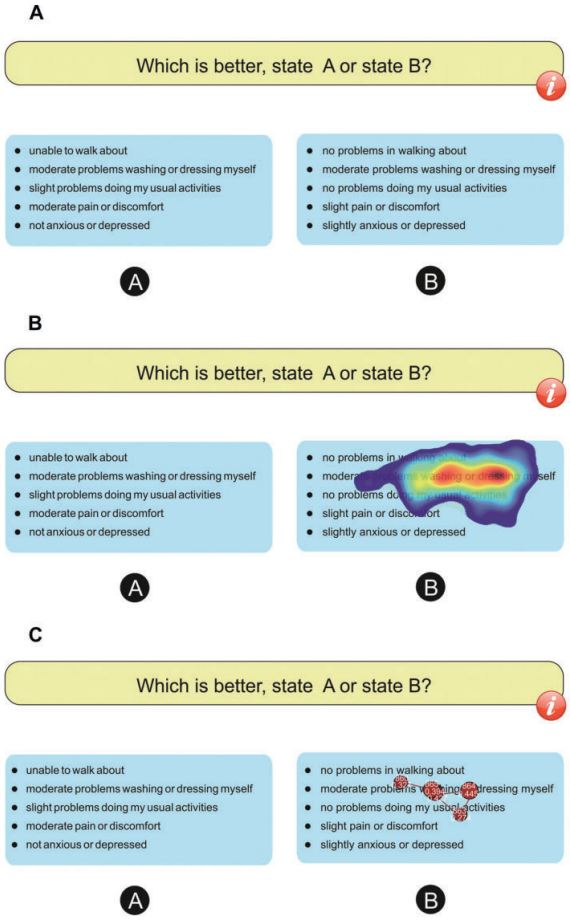

Fig. 1 Visualization options of: A) standard DC (paired comparison) response task based on EQ-5D-5L instrument, B) eye-tracking analysis based on heat map, C) eye-tracking analysis based on fixation path

Previous studies could not reach agreement on the optimal number of response tasks in DC studies [18-22]. In the health care settings, where the choices between various health states are cognitively demanding, the large number of tasks, such as 20 recommended by Johnson and Orme [23], would induce more fatigue. Therefore, we included the existing EQ-VT standard protocol format of one block consisting of seven tasks with two health-state descriptions each. The read out loud procedure which is part of the EQ-VT protocol was not used, as this might force the respondents to read all the information on the screen and might cause the feeling of being tracked. However, respondents were asked to complete six TTO tasks as warm up before proceeding with DC response tasks. This was done to familiarize the respondents with the health state description and the survey lay-out. The format was based on the EQ-5D-5L instrument (www.euroqol.org), which assesses functioning in five health attributes (domains), namely Mobility, Self-care, Usual Activity, Pain/Discomfort, and Anxiety/Depression, with five levels each (no problems, slight, moderate, severe, or extreme problems). The order of the health attributes (domains) was kept unchanged throughout the whole experiment. The respondents were expected to take into account the full description of the health states and to devote attention to all five attributes.

2.3 Eye-tracking analysis

The literature has shown that the fixation point of the eyes is linked to an individual's point of attention [24]. By implication, monitoring eye movements allows the researcher to capture which areas in the task with given information are attended and taken into consideration by that individual. Eye tracking is performed to capture the eye movements or the points of gaze thereby establishing the area on which the eye is fixated at a particular time while perusing the visual scene [25]. Generally, to see an object, the eye needs to fixate the gaze for a certain length of time, typically between 200ms and 600ms. The process of vision refers to scanning the object with rapid eye movements, called saccades [26]. While there are various techniques to track eye movements using special glasses or head-mounted displays, a video-based eye tracker was used in this study. It is less intrusive, more convenient and flexible than head-mounted systems, therefore, reducing the feeling of being tracked for respondents [27]. The principle behind the eye tracker is to record eye movements and map them to the computer screen for the analysis. The videos form the basis for an analysis of the fixation position, which entails detecting changes in the eye and pupil location, which lead to changes in the coordinates of the gaze.

To describe the process of paying attention, we investigated whether respondents consider the full set of information elements or whether they disregard particular elements in health-state descriptions. In case respondents fixate their eyes longer and more often on particular elements, we investigated what these elements are. In the current study, the videos of all respondents' eye movements were recorded and then analyzed using

the following features of the Gazepoint software analysis tools: heat maps and fixation paths in video format; areas-of-interest statistics; and graphs. Heat maps and fixation paths represent the overall pattern of the respondents' points or areas of attention: the areas of attention, duration of fixations, and direction of fixations (Figs 1b and 1c).

In addition, we constructed areas of interest to analyze whether respondents are disregarding particular attributes and whether left/right asymmetry of focusing on the health state descriptions exists [28]. The notion of areas of interest is grounded in the observation that some objects are more interesting and attract more attention, so the eyes fixate on these specific objects [29].

The areas of interest constructed for this study comprise the area of health state A and the area of health state B. For those two areas, we compared the statistics across the whole array of seven DC tasks. Then, to examine whether respondents are disregarding particular attributes, we established areas of interest for the position of each attribute and calculated the number of revisits made (Fig 2).

The number of revisits specifies the process of paying attention to the area followed by switching to another area and then moving back. Disregarding particular attributes was associated with attribute non-attendance, and defined as disregarding relevant information contained in one or more attributes [30, 31]. We defined that zero number of revisits (comparisons) implies that the respondents do not look at the attribute, indicating disregarding the attribute. In the event of disregarding the attribute, the number of revisits would be zero. Further statistical information of constructed areas of interest indicated which attributes were accorded a greater level of attention by the respondents.

Respondent fatigue is associated with reduced involvement indicated by the reduced time spent for task, as the respondent gets bored or tired [16, 17]. To investigate this issue, the average time spent per task was analyzed. In addition, the relationship between the average time spent per task and the health states comprising the choice pairs was

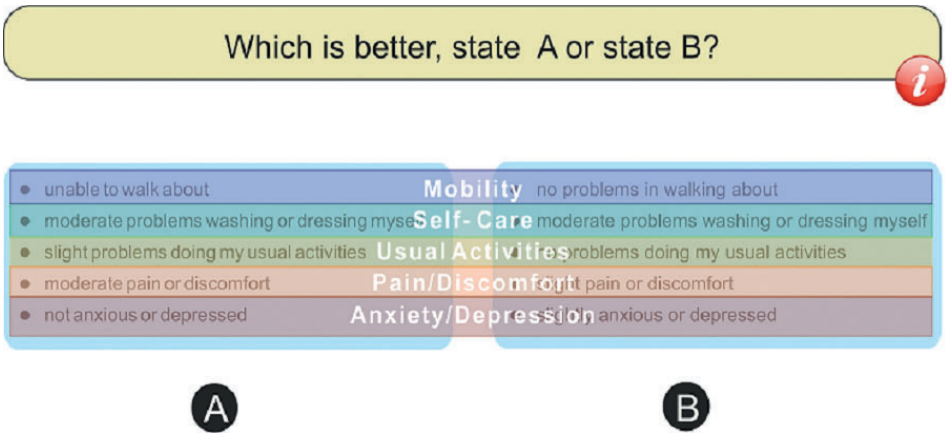

Fig. 2 Eye-tracking areas of interest for the five attributes (Mobility, Self-care, Usual Activities, Pain/Discomfort, Anxiety/Depression) of the two EQ-5D-5L health-state descriptions A and B.

investigated. We assumed that the complexity of the task is positively associated with the time spent on it. Finally, graphs were constructed to chart the process by which the DC response task was completed (Sigma Plot 13.0).

3. RESULTS

3.1 Data collection

The eye-tracking experiment was conducted during May and June 2016. First, data was collected from the six students, all of whom had completed the DC response tasks and the eye calibration procedure. The student sample consisted of three males and three females of age group 20-30 years old, without any eye problems or diseases and who did not wear glasses. Afterwards, data was gathered from seven members of the general public, of whom four were females and three males. The respondents represented age groups 30-40, 40-50, and 50-60 years old. Of the latter sample, four persons did not wear glasses during the experiment whereas three did. Due to the poor performance of the device for respondents wearing glasses - it gave imprecise locations for the eye fixation - the results for three respondents from the general public were excluded from the analysis.

3.2 Disregarding particular attributes

The revisit frequency statistics showed that none of the attributes had been overlooked, according to our definition of disregarding the attribute when the number of revisits is zero (Fig 3). Moreover, the average number of revisits differed among the respondents, indicating a divergence in the intensities of attention. Three respondents revisited Anxiety/Depression much less than any other attribute, while two respondents revisited Mobility less than the other attributes. Mobility (top) and Anxiety/Depression (bottom) were in general slightly less frequently attended than the other attributes (Fig 3).

3.3 Left/right asymmetry of focusing

The analysis revealed a tendency to focus the eye slightly longer on health-state descriptions presented on the left side of the slide. We calculated the duration and the number of fixations on the left side of the choice task, depicting health state A, versus the right side, depicting health state B (Figs 4a and 4b).

Both indicators (duration and number of fixations) showed similar results: the cases with longer duration had the higher number of fixations, and vice versa. All respondents fixated slightly longer and more often to the left-side health state. However, three respondents fixated their eyes longer on the right-side health state than on the left-side health state. In general, the following differences across all respondents were observed: 1-10 seconds longer fixation time on the left side over the right side (for the duration of

fixation), and 3-7 fixations more for the left-side health state over the right-side (for the number of fixations).

3.4 Attention to specific attributes

We assumed that the duration and the number of fixations are feasible indicators of attention paid to a specific attribute. It may be noted that for the attributes with a longer duration of fixation, the number of fixations was also generally larger (Figs 5a, 5b). A longer duration of fixation on an attribute is taken to be indicative of the importance of this attribute for making a choice in the DC response task. The frequency statistics of fixation on each attribute (in seconds) showed differences in attention paid to the attributes (Fig 5a). Specifically, Self-care and Usual Activities had a longer duration of fixation. The attributes Mobility and Anxiety/Depression attracted less attention, although two respondents had fixated their eyes longer on Anxiety/Depression. Importantly, Pain/Discomfort was slightly less attended than Self-care or Usual Activities, although one respondent fixated the eyes mostly on Pain/Discomfort.

3.5 Respondent fatigue

No increasing or decreasing trend in the average amount of time spent per task was observed (Fig 6).

Less time was spent for the easier choice tasks (severe state versus mild state) than for more complex tasks entailing a choice between pairs with similar levels of severity. For example, in the second DC response task, a very mild state (11112) is compared with a health state consisting of moderate and severe problems (33243). Therefore, the time spent decreased in comparison with the first or third tasks, where severe health states are presented. However, the results for the last two tasks containing severe health states showed a decrease in time spent per task.

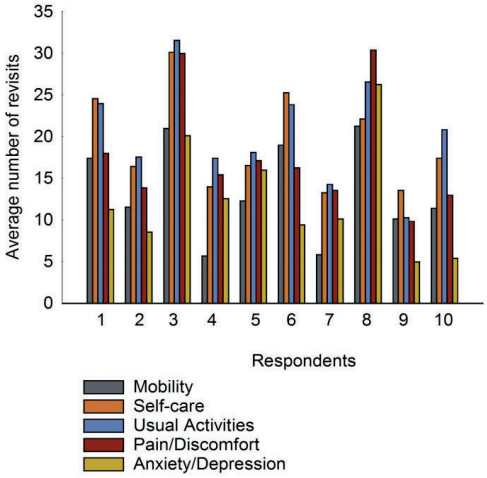

Fig. 3 Average number of revisits on the five EQ-5D-5L health attributes per DC response task

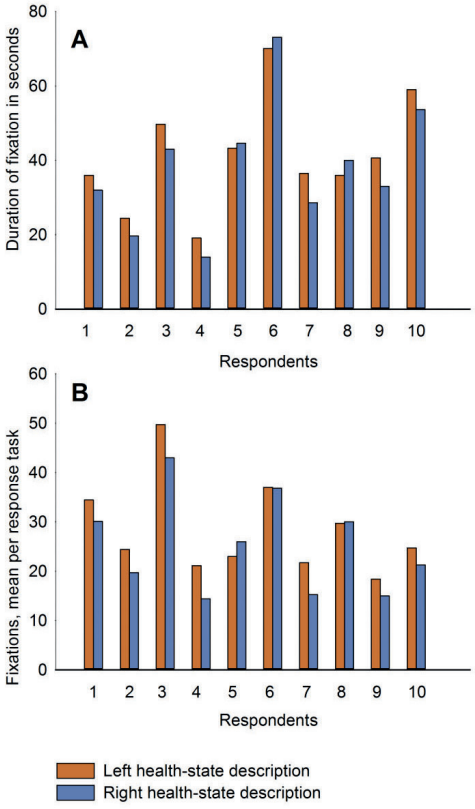

Fig. 4 Comparison of the two health states for the left and right-side versions of the DC task (health state A and B respectively): A) duration of fixation in seconds; B) mean fixations per response task

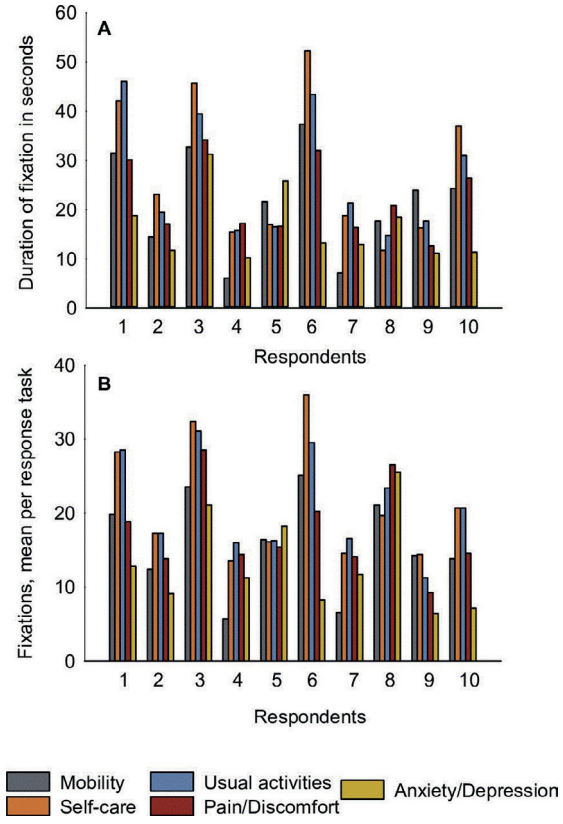

Fig. 5 Duration of fixation in seconds (A), and mean number of fixations per DC response task (B) for five EQ-5D-5L health attributes

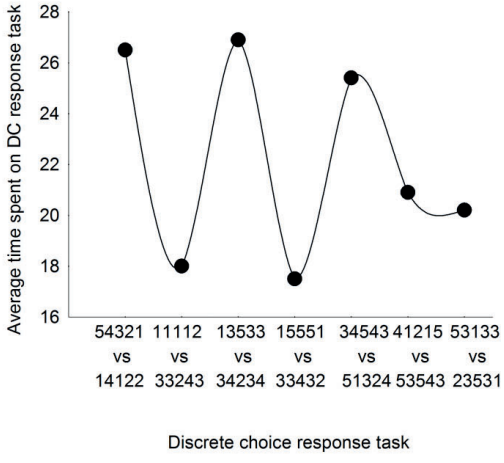

Fig. 6 Relationship between average time spent per DC response task and health state pair

4. DISCUSSION

The aim of this study was to detect whether respondents pay attention to all elements of the health-state descriptions in discrete choice (DC) response tasks. To summarize, we found that the respondents pay attention to all elements of the health-state descriptions in the tasks. However, this explorative eye-tracking study revealed differences in attendance for different EQ-5D-5L attributes and a slight longer fixation time on the health-state description on the left side. Respondent fatigue was not demonstrated, however, time spent per task seem to be influenced by the complexity of the task.

The attributes Self-care and Usual Activities (2nd and 3rd) were visited most frequently. Their attraction may be explained by their central position on the screen and by the relative length of the description for these attributes. Another possible explanation is that higher severity levels of an attribute attract more attention compared to attributes of low severity. According to Orquin et al. [14], respondents tend to gaze at information with a greater utility or importance to their decision. In our case, severe levels of attributes represent shifts away from good health, which may be considered most important for making a choice between two health-state descriptions. In a study by Spinks and Mortimer [32] using eye tracking, non-attendance (or disregarding relevant attributes) was found to be associated with the complexity of the task as well as with ordering effects, time limits, and survey-specific effects. The present study used a fixed ordering of the attributes, allowing the respondents to get used to the sequence and develop their own strategy for making their choice. In an earlier study, it was observed that the alternative (i.e., health description) examined first and the centrally positioned alternative receive more attention [13]. In the possible alternative versions of our survey the random ordering of the attributes could be implemented. In this case the higher positioned attributes Pain/Discomfort and Anxiety/Depression in the health description, are more likely to be attended than in the current version of our study, where these attributes are positioned lower.

The analysis of the asymmetry of attendance process for the health-state description showed a slightly longer fixation time on the left side. The tendency to fixate longer on the left side than on the right side was studied earlier and defined as left-to-right bias, which could be due to the regularity of the direction in which individuals read text. An earlier study [33] confirmed left-to-right bias in a sample of English-speaking participants who read in a predominantly left-to-right manner, but the opposite bias was observed in an Arabic sample, who read text from right to left. However, in our study a difference in the range of 1-10 seconds longer fixation time on the left side over the right side in the duration of fixation throughout the whole survey was inconclusive. A difference in the range of 3-7 fixations per DC response task was not large enough to make a conclusive judgment. By contrast, in two cases in our study attention was mostly paid to the right side. A study by Kinsbourne [34] suggested that there is asymmetry in information perception, implying that eye movements can be biased, depending on whether the

process is controlled by the left or the right hemisphere. In a subsequent study [35] Kinsbourne found that right-handed people tend to pay more attention to the left side while the opposite holds for left-handed people. These findings could explain the slight preference we observed for alternatives presented on the left side, considering that most of the population is right-handed.

We found no evidence of respondent fatigue after completing 13 tasks, taking into account that before the DC response tasks the respondents had performed six TTO tasks (not part of this study). Previous studies were inconclusive about the number of tasks that should be included to avoid respondent fatigue. According to Craig et al. [36], neither randomization nor fewer response tasks would affect response precision, a statement that contradicts prior findings [37-39]. On the other hand, Savage et al. [40] found that respondents suffered fatigue in online surveys with repeated tasks. Finally, Carlsson and Martinsson [41] suggested that the sequence of the tasks does not induce fatigue, as the respondents are capable of answering many choice sets. However, it needs to be remarked that the time spent per DC task in our study was associated with the difficulty of the task, such as relative similarity of health-states under comparison. These findings are in line with the results of the earlier study [14] suggesting that the information sampling needed to reach a decision increases as options become more similar.

This explorative study may be considered a first step towards the application of eye tracking to detect possible fixation strategies that respondents may use in their processes of paying attention to the information cues during the completion of DC response tasks for health-state evaluation. For the DC response task in the context of health states, this study has shown only limited differences in attendance. However, further investigation is warranted, requiring various designs for the more complex valuation techniques or for DC studies with multiplex scenarios (combining: health attributes, life time duration, death). For more complex situations, changing the attribute order and including additional visual cues to enhance the understanding of the health-state description may reveal whether attendance depends on the location of an attribute description rather than (or as well as) on its content.

Some limitations of this study should be mentioned. First, the sample is not representative and the sample size is small. Due to the exploratory nature of this study, we considered ten respondents sufficient to reveal important differences in the attendance process. Moreover, the eye-tracker equipment could not capture precisely the eye movements of people wearing glasses. The imprecision was due to the reflection of the lenses, irrespective of the presence or absence of daylight, the positioning of the respondent, the type of room for the experiment, or the screen type. This is considered a limitation because a large portion of the general public wear glasses and they would find it inconvenient or impossible to read the tasks without wearing glasses. Additionally, the eye tracker has a certain level of precision (0.5-1 degrees) which was found to be critical in the designation of small areas of interest as description statements for attributes.

Therefore, further studies would require more sophisticated equipment to tackle these issues.

A central assumption in this study is that information that is seen is also cognitively processed. When using fixations and saccades to quantify the amount of attention as in the present study, researchers base their analysis on this assumption. Although the assumption is generally validated, there are also empirical findings indicating that eye movements do not necessarily reflect cognitive processes in certain decision contexts [42]. Normally, attention is paid in the focus area (fovea), but it is also possible that humans pay attention to objects currently outside this area [14]. In the current study, it is possible that peripheral viewing would allow more experienced respondents to pay attention to objects which are outside their focus area. Therefore, caution in the definition of paying attention and understanding or processing information is needed.

The current study was based on fixations (number and length). Eye-tracking capabilities may also include information about pupil dilations and the number of eye blinks. The investigation of pupil dilations can be used in the future research for the tasks with changing complexities because dilations can be a consistent index of cognitive load [43]. Furthermore, eye blinks can be of interest to describe the processing flow and indicate triggering and cognitive shifts.

In conclusion, the current study investigated the process of attendance to various information cues presented in response tasks. Overall, respondents tend to pay attention to the full description and do not use shortcuts or disregard particular information elements.

5. REFERENCES

1. Levine, S. The meaning of health, illness, and quality of life. Geggenmoose-Holzman I, Brenner H, Flick U, editors. Quality of life and health: concepts, methods and applications. Berlin: Blackwell Wissenschaft; 1995.
2. Hamming JF, De Vries J. Measuring quality of life. Br J Surg. 2007; 94: 923–924.
3. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technol Assess. 1999; 3(9):57-76.
4. Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. J Health Econ. 1997; 16: 155-75.
5. Doctor JN, Bleichrodt H, Lin HJ. Health Utility Bias: A Systematic Review and Meta-Analytic Evaluation. Med Dec Making. 2010; 58-67.
6. van Osch SMC, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting Biases in Standard Gamble and Time Tradeoff Utilities. Med Decis Making. 2004; 511-517.
7. Van der Pol M, Roux L. Time preference bias in time trade-off. Eur J Health Econ. 2005; 107-11.
8. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. J Health Econ. 2002; 11(5):447-56.
9. McFadden D. Modeling the choice of residential location. In: Karlqvist A, Lundqvist L, Snickars F, Weibull J. Spatial Interaction Theory and Planning Models.eds. North-Holland, Amsterdam; 1978.
10. Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. Value Health. 2010; 13: 1005-13.
11. Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: Background, theories, assumptions and applications. Expert Rev Pharmacoecon Outcomes Res. 2013; 13(1): 93–108.
12. Lohse GL, Johnson EJ. A comparison of two process tracing methods for choice tasks. Organ Behav Hum Decis Process. 1996; 68 (1): 28–43.
13. Meißner M, Musalem A, Huber J. Eye-tracking reveals a process of conjoint choice that is quick, efficient and largely free from contextual biases. J. Mark. Res. 2016; 53 (1): 1-17.
14. Orquin JL, Mueller Loose S. Attention and choice: A review on eye movements in decision making. Acta Psychologica. 2013; 144: 190–206.
15. Pernice K, Nielsen J. Eyetracking Methodology. How to Conduct and Evaluate Usability Studies Using Eyetracking. Nielsen Norman Group. 2009; 19-52.
16. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. Proceedings of ACM INTERCHI'93 Conference. 1993; 206-13.
17. Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37: 53–72.
18. Caussade S, Ortúzar JD, Rizzi L, Hensher DA. Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates. Transp Res B. 2005; 39: 621-640.
19. Brazell JD, Louviere JJ. Helping, Learning, and Fatigue: An Empirical Investigation of Length Effects in Conjoint Choice Studies, Department of Marketing, the University of Sydney; 1996.
20. Phillips KA, Johnson FR, Maddala T. Measuring What People Value: A Comparison of 'Attitude' and 'Preference' Surveys. Health Serv Res. 2002; 37: 1659-1679.

21. Bech M, Kjaer T, Lauridsen J. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. J Health Econ. 2011; 20: 273–286.
22. Hess S, Hensher D, Daly AJ. Not bored yet – revisiting respondent fatigue in stated choice experiments. Transp Res: Policy and Practice. 2012; 46(3):626-644.
23. Johnson RM, Orme BK. How many questions should you ask in choice-based conjoint studies?" Research paper series 153. Washington, DC: Sawtooth Software; 1996.
24. Hoffman JE & Subramaniam B. The role of visual attention in saccadic eye movements. Percept Psychophys. 1995; 57(6): 787-795. doi:10.3758/BF03206794
25. Majaranta P, Bulling A. Eye Tracking and Eye-Based Human–Computer Interaction. In Advances in Physiological Computing, Human–Computer Interaction Series, S. H. Fairclough and K. Gilleade, (eds.) Springer-Verlag London; 2014.
26. Jacob RJK. Eye tracking in advanced interface design. In: Barfield W, Furness TA, (eds.).Virtual environments and advanced interface design. Oxford University Press, New York; 1995.
27. Chen YS, Su CH, Chen JH, Chen CS, Hung YP, et al. Video-based eye tracking for autostereoscopic displays. Opt Eng. 2001; 40(12): 2726-2734. http://dx.doi.org/10.1117/1.1416130.
28. H. Weber, B. Fischer. Gap duration and location of attention focus modulate the occurrence of left/right asymmetries in the saccadic reaction times of human subjects. Vision Res.1995; 35 (7): 987-998. https://doi.org/10.1016/0042-6989(94)00186-P.
29. Orquin JL, Ashby NJS, Clarke ADF. Areas of Interest as a Signal Detection Problem in Behavioral Eye-Tracking Research. J Behav Dec Making. 2016; 29: 103–15.
30. Hensher DA. Attribute processing, heuristics and preference construction in choice analysis. In State-of Art and State of Practice in Choice Modelling, Hess S, Daly A, (eds.) 2010. Emerald Press: U.K.
31. Lagarde M. Investigating attribute non-attendance and its consequences in choice experiments with latent class models. J Health Econ. 2013; 22: 554–567.
32. Spinks J, Mortimer D. Lost in the crowd? Using eye-tracking to investigate the effect of complexity on attribute non-attendance in discrete choice experiments. BMC Med Inform Decis Mak. 2016; 16:14.
33. Spalek TM & Hammad S. The left-to-right bias in inhibition of return is due to the direction of reading. Psychol. Sci. 2005; 16 (1):15-18.
34. Kinsbourne M. The cerebral basis of lateral asymmetries in attention. In Acta Psychologica 33 Attention and Performance III, Sanders AF, (eds.) North-Holland Publishing Company, Amsterdam; 1970.
35. Kinsbourne M. Eye and Head Turning Indicates Cerebral Lateralization. Science. 1972; 176: 539-541.
36. Craig BM, Runge SK, Rand-Hendriksen K, Ramos-Goñi JM, Oppe M. Learning and Satisficing: An Analysis of Sequence Effects in Health Valuation. Value Health. 2015; 18:217 – 223.
37. Holmes TP, Boyle K. Dynamic learning and context-dependence in sequential, attribute-based stated preference valuation questions. Land Econ. 2005; 81:114–126
38. Hensher DA, Stopher PR, Louviere JJ. An exploratory analysis of the effect of number of choice sets in designed choice experiments: an airline choice application. J Air Transp Manag. 2001; 7:373–379.

39. Chung C, Boyer T, Han S. How Many Choice Sets and Alternatives are Optimal? Consistency in Choice Experiments. Agribusiness. 2011; 27: 114–25.

40. Savage SJ, Waldman DM. Learning and fatigue during choice experiments: a comparison of online and mail survey modes. J Appl Econ. 2008; 23: 351–371.

41. Carlsson F, Martinsson P. How much is too much? Environ Resour Econ 2008; 40: 165–76.

42. Anderson JR, Bothell D, Douglass S. Eye movements do not reflect retrieval processes: limits of the eye-mind hypothesis. Psychol. Sci. 2004; 15 (4): 225-231.

43. Just MA, Carpenter PA. The intensity dimension of thought: pupillometric indices of sentence processing. Can J Exp Psychol. 1993; 47 (2):310–39.

GENERAL DISCUSSION

GENERAL DISCUSSION

Importance of the studies and overview of the issues

Over the past decades, as advances in medical treatments have improved survival and reduced key morbidities, perceived health status or health-related quality of life (HRQoL) assessments are becoming more and more relevant. HRQoL is a measure of perceived health status consisting of physical, mental, and social domains [1]. Regulatory bodies such as the Food and Drug Administration (FDA) [2] and the National Institute for Health and Care Excellence (NICE) [3] actively encourage measuring patient-reported HRQoL in addition to traditional clinical assessments. The public and patients are asked to make value judgments about different aspects of health to obtain a subjective yet numerical expression of the quality of a certain health state.

HRQoL instruments can be developed on the basis of various measurement frameworks. However, when comparing HRQoL across different populations, conducting disease modeling studies, and performing economic evaluations of various healthcare interventions, it is more reasonable to use preference-based instruments. These differ from other measures in that, by expressing the evaluation in a single metric score, they explicitly incorporate weights that reflect the importance attached to specific health aspects. A single metric score represents the quality of a health state holistically and is referred to as a health-state value.

The aim of this thesis was to investigate the specific problems associated with preference-based measures of health states and with the methodology used to derive health-state values. Health-state values can be at a cardinal or an interval measurement level [4]. If these values are transformed or normalized on a scale ranging from death to full health, they become utilities. Such utilities are often used in computing quality-adjusted life years, a key notion in cost-effectiveness analysis. Throughout this thesis we applied discrete choice modeling, which produces values, which cannot be formally referred to as utilities, but for our methodological investigations the use of values is sufficient.

Since the research underlying the present thesis required a large number of respondents, we needed to find a simple instrument enabling a self-completion format. Therefore, we selected the EQ-5D, a relatively simple and widely used generic preference-based instrument. For example, the guidelines for pharmacoeconomic research in the UK and in the Netherlands recommend the EQ-5D instrument for health-state evaluations [5, 6]. In the EQ-5D the description of any health-state can be presented by means of five health domains (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), where each domain has a limited number of levels. Different health domains (also called attributes) are weighted on the basis of comparisons the respondents have to make. For example, they may be asked to choose between health-state descriptions. Their choices allow the researchers to determine the relative importance of these attributes and of the levels for each attribute [7, 8].

The valuation framework of discrete choice modeling, which was selected for this thesis, is consistent with the random utility model in economic theory [9]. Now it is used to augment economic evaluations in the realm of healthcare with information about the value of non-health outcomes such as waiting time, location of treatment, and type of care [10]. Moreover, this technique has been used to elicit personal and societal preferences in health-valuation studies [10]. Discrete choice models started to attract attention as an alternative to the valuation frameworks of time trade-off and standard gamble; these latter techniques were complex and prone to bias (time preference, states worse than dead, risk averse, scale compatibility) [11-15]. A promising feature of discrete choice models is that the derived values only relate to the attractiveness of a health state. They are not expressed in trade-offs between improved health and something else, as in time trade-off and standard gamble. Discrete choice is considered a relatively easy task for the respondents since it mimics individual everyday choices: 'Which of the available options is more preferable?' In comparison, conventional valuation techniques, such as time trade-off and standard gamble, are based on an iterative procedure, whereby a choice is made between each of the presented options till the point of indifference is reached. This task is challenging for the respondents and has led to biases. By contrast, respondents can tackle up to 20 discrete choice tasks without feeling fatigued. This enables the researchers to get enough information to specify more complex health values. Discrete choice tasks can be conducted online, thereby cutting out the interviewer (who would be needed for the time trade-off or standard gamble) and eliminating possible interviewer effects on the responses [16, 17]. However, a known limitation of discrete choice models is their inability to produce values in the form of absolute numbers. All values in discrete choice models are relative and are located on a latent scale from 'best' to 'worst' states. That limitation can be attributed to the location of 'dead', which is unknown since that option was not included as a choice option in the response tasks. Therefore, one of the main problems with discrete choice models is how to normalize the scale to dead– full health (0.0 – 1.0). To solve this problem and enable utility calculations, the design should include a task extension or additional tasks. In this thesis, we did not normalize the values on dead and full health to generate utilities. Instead, we focused on other aspects of health-state measurement. Specifically, we devoted our attention to four issues: the content and description of health states; problems with preference-based estimation (interactions); whose responses should be considered (those of the general population or of patients); and the respondents' attention to the discrete choice tasks.

Issue 1: Content and description of health states

Various generic instruments that have been developed to provide values or utilities of health states define and describe the same health states in different ways. Consider, for example, the relatively simple instrument developed by the EuroQol Group the EQ-5D [18, 19]. In the

standard version (EQ-5D-3L) each of the attributes can take on three levels [20]. However, it was presumed to have specific drawbacks - a restricted discriminatory power, lack of sensitivity to smaller changes in health states, a substantial proportion of respondents report themselves as being in the best health state (ceiling effect) - which prompted an update of the instrument [21, 22]. For the new version, the EQ-5D-5L, the number of levels used to classify health states was increased from three to five, and the phrasing of health-state descriptions was changed. According to several earlier studies, the differences between the levels in the EQ-5D-5L are subtle and may be hard to distinguish, which might have caused some of the language versions (notably the English) to show inconsistencies at the upper or lower levels of health attributes [23-25]. Such inconsistencies would affect the validity of the estimated values. The current Dutch study (Chapter 1) used discrete choice modeling, thereby deviating from earlier studies [23-25], and found no inconsistencies for either version of the EQ-5D. However, the overall weights for the attributes were different in the two versions. In the EQ-5D-5L the highest weight was attributed to anxiety/depression, while in the EQ-5D-3L the highest weight was attributed to mobility. A change made in the wording of the description of the mobility attribute from 'confined to bed' to 'unable to walk' is a possible explanation for the shift in the level of importance. In that light, researchers should be cautious about describing health states with existing instruments or revising the descriptions. Apparently even small differences in wording could affect individual responses and thereby the elicited values. Another concern is how comprehensively the health state needs to be described. Insufficient detail could lead the respondents to start guessing, thus distorting the actual description of the health state in question. On the other hand, excessively detailed descriptions would overload the respondents with information to be processed, thereby creating fatigue and leaving more room for random responses.

Issue 2: Interactions between attributes

Attention was also given to the inclusion of interactions between distinct health attributes. Not only did we find that such interactions exist, but we showed that the combined effect of two separate health attributes is stronger than the sum of their individual effects. Most models are main-effects models, which take into account the individual effects but not the interactions of distinct health attributes. However, we suggest that interactions of health attributes should be taken into account. In that regard, future studies might test whether specific interactions exist between health attributes of interest. For example, a researcher might seek to verify that the most salient interactions found in the present study would again appear with a different version of an instrument and a different setting. The results of the current study are applicable only to the original version of the generic preference-based EQ-5D instrument, where the health-state descriptions are based on health attributes with three levels each. It would be important to find out whether similar results can be achieved with other preference-based instruments, such as HUI or SF-6D, because the set of attributes used for the construction of instruments can influence

the strength of the interactions between these attributes. However, these instruments use larger sets of attributes and levels, which makes it harder to evaluate all possible interaction terms. In this case, we have to rely on theoretical knowledge and the literature to test the significance of specific interactions between the health attributes of interest.

Issue 3: Whose responses?

Conventionally, the health-state values used in economic evaluations are derived from a representative community sample, a convention based on several arguments [26, 27]. One is that, as taxpayers, these persons are deemed to represent the general population [28, 29]. Another is the 'veil of ignorance' argument [28], whereby the general population is presumed to have never experienced the impaired health states under evaluation and to be blind to its own self-interest. Accordingly, representatives of the general population would embody principles of justice and equity and, thereby, ensure a fair distribution of resources. However, a community sample consists mainly of healthy or relatively healthy persons, who may be inadequately informed or have insufficient imagination to make an appropriate judgment about the impact of (severe) health states. Instead, judgments made by patients are put into the spotlight, assuming that people who have direct experience with impaired health will provide more reliable and informative health-state valuations [30, 31]. A lack of consensus on whose values to use motivated the investigation of whether any differences between the public and patients' judgments actually exist. In that vein, an important aspect of this thesis was to compare different population samples. A comparison was made between healthy respondents and patients regarding health-state valuations, and a comparison was made between the general public and patients regarding new medical treatments (Chapters 3 and 4). We did not find large differences between the health values elicited from the general population and patients, but we did find differences between those of patients and healthy respondents. The values elicited from the general population can be quite similar to the values of healthy respondents (in case the majority of the population is healthy) or to those of patients (in case the majority of the population has some experience of disease). These results suggest that the sample of a representative general population should be constructed with some caution. While sampling based on demographic characteristics is important, it is also important to ensure an equal representation of the proportion of respondents with and without specific diseases. The reason for such caution is to avoid that overall population values reflect only healthy individuals or only patients. However, it may happen that patients who have certain diagnosed diseases might not have experienced some of the health states they are asked to evaluate and therefore might not be able to give informed judgments based on actual experience. By the same logic, healthy respondents could have been patients in the past and, thus, have experienced the health states under evaluation, enabling them to give informed answers based on experience. This discrepancy raises the issue of experience-based values. In the future, researchers could consider asking respondents to value health states that lie nearby their own (current or past) health

status. In that way, the health states would not be completely hypothetical. Unfortunately, we did not ask the respondents in our studies to specify whether they had experienced the health states under evaluation, but it would be an interesting area for future study.

Issue 4: Attention of respondents to the discrete choice tasks

The last issue raised in this thesis was prompted by a basic assumption in the preference-based measurement framework. It is generally assumed that respondents pay equal attention to all components of information (e.g. left-side versus right-side alternatives, specific attributes used in the description) presented in the response task. These tasks require the respondent to make considered choices among two or more scenarios described by specific attributes with certain levels. Using an eye-tracking device, we investigated whether the respondents were paying attention to all of the information that was presented in the discrete choice tasks. Disregarding specific attributes would lead to an incomplete understanding of the health-state descriptions and, thus, to biased results with the statistical models applied in health evaluations. However, the eye-tracking study revealed that the respondents did pay attention to all of the information elements and were not fatigued after completion of the task. It should be mentioned that we used a specific layout (EQ-VT, a standardized valuation study protocol) for the study. The effect of applying eye-tracking to different versions or different layouts, such as a changed ordering of the attributes, different color schemes, different placement of information cues on the screen, could be an interesting topic for future research. The results would give insight into the optimal layout and appearance of the health valuation instrument, allowing researchers to simplify it and make it more attractive for the respondents. The effectiveness of discrete choice studies depends strongly on the survey design, and not only on the experimental design but also on the layout design. Careful selection of health-state descriptions and survey elements (phrasing, instructions, ease of navigation, information notes, and visual cues) are required to conduct a study. When these standards are met, the respondents are more likely to take account of the full range of information presented and make more informed judgments. Moreover, the problem of fatigue would be reduced by making the design simple and attractive.

To sum up, this thesis has provided evidence in support of using discrete choice modeling for health-state measurements. The present study has drawn attention to some important issues: the content and description of health states; interactions between health attributes; the construction of samples to derive the health values; and the importance of the survey design to enhance the respondents' attention to the response tasks. We pointed out that preference-based health-state measurement is associated with several methodological drawbacks that might warrant attention in future research. First, a simple main-effect model for health-state measurement may not be sufficiently accurate to produce credible health-state values. Therefore, interaction terms would need to be tested and studied carefully. Second, researchers would need to be prudent when

developing or modifying a preference-based instrument. Even small differences in the phrasing or the valuation technique in combination with particular statistical models may affect the expected results and elicited values. Third, the judgments of healthy people may differ from those of patients who are actually experiencing health limitations. It would be advisable to use values based on assessments by patients instead, as patients are likely to be more adequately informed than healthy people or more adept at imagining certain health states. Accordingly, patients may be better motivated to make an informed judgment about the impact on perceived health of such states. However, if the patient community wants to have a central role in defining value, robust models would be needed to incorporate the patient voice in a value assessment that is free from adaptation and other biases. Fourth, it is important to invest in the design and layout of preference-based instruments in general and discrete choice tasks in particular, with the objective of making the task attractive to the respondents without causing fatigue. We believe that a logical, simple, and reliable measurement model is needed for the measurement of a subjective phenomenon such as health status. The discrete choice model and various related choice models are probably qualified candidates.

REFERENCES

1. World Health Organization. The first ten years of the World Health Organization. Geneva: World Health Organization, 1958
2. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH). 2009
3. Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing Health-Related Quality of Life: An EQ-5D-5L value set for England. Health Econ. 2017; 1-16.
4. Krabbe PFM. The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. San Diego, USA: Elsevier/Academic Press; 2016
5. Guidelines for pharmacoeconomic research, updated version. College voor zorgverzekeringen, Diemen March 2006
6. Viney R, Norman R, Brazier J, et al. An Australian discrete choice experiment to value EQ-5D health states. J Health Econ 2014; 23:729-742
7. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: Health utility index. Pharmacoecon 1995; 7: 503–520
8. Fischer GW. Utility models for multiple objective decisions: do they accurately represent human preferences? Decis Sci 1979; 10: 451–479
9. Arons AMM, Krabbe PFM. Probabilistic choice models in health-state valuation research: background, theory, assumptions and relationships. Expert Rev of Pharmacoecon Outcomes Res. 2013;13: 93–108.
10. Stolk EA, Oppe M, Scalone L, Krabbe PFM. 2010. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. Value Health. 2010; 13, 1005-1013.
11. Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. J Health Econ. 1997; 16: 155-75.
12. Doctor JN, Bleichrodt H, Lin HJ. Health Utility Bias: A Systematic Review and Meta-Analytic Evaluation. Med Dec Making. 2010; 58-67.
13. van Osch SMC, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting Biases in Standard Gamble and Time Tradeoff Utilities. Med Decis Making. 2004; 511-517.
14. Van der Pol M, Roux L. Time preference bias in time trade-off. Eur J Health Econ. 2005; 107-11.
15. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. J Health Econ. 2002; 11(5):447-56.
16. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics 2008; 26: 661-77.
17. Krabbe PFM, Devlin NJ, Stolk EA, Shah KK, Oppe M, van Hout B, Quik EH, Pickard AS, Xie F. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014; 52(11): 935-943.

18. Hurst N, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in people with rheumatoid arthritis: validity, responsiveness and reliability of the EuroQoL (EQ-5D). Br. J. Rheumatol. 1997; 36: 551–559.
19. Rabin R, Charro de F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001; 33:337-343.
20. Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37 (1), 53–72.
21. Badia X, Herdman M, Kind P: The influence of ill-health experience on the valuation of health. Pharmacoecon. 1998, 13:687-696.
22. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D-3L and SF-6D across seven patient groups. J Health Econ. 2004; 13(9):873–84.
23. Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PFM. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best–worst scaling? Eur J Health Econ. 2014. doi: 10.1007/s10198-013-0474-3.
24. Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? Qual Life Res. 2015; 24(7): 1759–1765.
25. Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016; 19 (4): 343-352.
26. Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the economic evaluation of health care programmes. Fourth ed. Oxford University Press; 2015.
27. Neumann PJ, Ganiats TG, Russell LB, et al. eds. Cost-Effectiveness in Health and Medicine. Oxford University Press; 2016.
28. Gandjour A. Theoretical foundation of patient v. population preferences in calculating QALYs. Med Decis Making 2010; 30 (4): 57-63.
29. Rand-Hendriksen K, Augestad L, Kristiansen IS, et al. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. Qual Life Res 2012; 21:1005–1012.
30. Devlin NJ, Appleby J. Getting the Most out of PROMs. Kings Fund, London; 2010.
31. Nilsson E, Orwelius L, Kristenson M. Patient-reported outcomes in the Swedish National Quality Registers. J Intern Med 2016; 279(2): 141-53.

SUMMARY

Assessment of perceived health status or health-related quality of life (HRQoL) is becoming more and more relevant. HRQoL is a measure of perceived health status consisting of physical, mental, and social domains. There are different approaches to HRQoL measurement. In this thesis we focus on the preference-based framework, which captures a person's overall health condition or health status in a single figure. Such single figures are referred to as values, as they explicitly incorporate weights reflecting the importance attached to specific health aspects.

The preference-based framework is used in several instruments (e.g., EQ-5D, HUI-3, SF-6D, AQol). For this thesis we used the EQ-5D instrument, which comprises five domains: mobility, self-care, usual activity, pain/discomfort, and anxiety/depression. Some methodological drawbacks of its three-level version (EQ-5D-3L) prompted development of a new format (EQ-5D-5L). There is no clear evidence that the new format outperforms the standard version, however.

The objective of the **first** chapter was to make a head-to-head comparison of the EQ-5D-3L and EQ-5D-5L in a discrete-choice model setting. The discrete-choice valuation framework is grounded in modern measurement theory and is consistent with the random utility model in economic theory. The discrete-choice technique requires participants to make choices among two or more presented scenarios (choice tasks) described by means of specific domains (also called attributes) with certain levels. Special attention was paid to the consistency and logical ordering of coefficients for the attribute levels, as well as to differences in health-state values. Values derived from a large representative sample of the 3,698 members of the Dutch population using the same measurement framework for both versions showed slight differences between the EQ-5D-3L and EQ-5D-5L. No inconsistencies or illogical ordering of level coefficients were observed in either version. We reason that even small differences in the phrasing (language) of the descriptive system or the use of another measurement method could produce differences in values between these two versions.

One of the common features among preference-based instruments (e.g., SF-6D, EQ-5D, AQol) is that they use statistical value functions comprising only the main effects for the separate health domains in their measurement procedures. Such functions rely on the simplifying assumption that the overall effect of all HRQoL domains equals the sum of the components' effects (i.e., reduction in perceived health status may intensify if two different health problems interact). However, health domains are often related to and depend upon each other. In earlier studies (Feeny et al., 2002; Rowen et al., 2009) it was demonstrated that the effect of health-state domains is not simply additive, and that interactions may be important. In the development of the EQ-5D instrument the N3 term was used in several models to determine the presence of level-3 severity in any of the domains of the health state. However, no proof was found of either intuitive appeal or an improvement of model fit after inclusion of this omnibus interaction term. Therefore, the

objective of the **second** chapter was to investigate whether inclusion of second-order interactions in the EQ-5D-3L value function would result in better fit and lead to different health-state values than a model with main effects only. Using an efficient design, 400 pairs of EQ-5D-3L health states were generated in a pair-wise choice format. We analyzed responses of 4,000 persons from the general population using a conditional logit model and tested its goodness-of-fit. Overall, for the EQ-5D-3L, a value function based on interactions produced systematically lower values than a main-effects model, meaning that the effects of two or more health problems combined was stronger than the sum of the individual main effects.

Typically, the health-state values used in economic evaluations are derived from a representative community sample comprised mainly of healthy or relatively healthy persons on the grounds that, as taxpayers, they represent the general population. Recently, investigators in many countries have turned to deriving values from patient preferences, reasoning that patients are better informed about or more able to imagine certain health states and therefore better able to evaluate these. The **third** chapter uncovers differences between contrasting samples (people with disease experience versus currently healthy respondents) in the importance they assign to various EQ-5D-5L health states and in the underlying health-state values. The responses of 3,068 respondents were analyzed and the values assigned to each EQ-5D-5L health state were compared. Differences in appraisal of health states were found between individuals who had experienced disease and those who had not, resulting in dissimilarity in health-state values between healthy individuals and patients. This study emphasizes the importance of eliciting values from respondents who have experienced certain health states or diseases.

The **fourth** chapter examines differences between the two samples (general population and patients), although in a different setting than the EQ-5D instrument. The objective of this chapter was to determine the importance of certain criteria regarding new treatments and explore whether there are differences in preference for these criteria between the general population and patients. This issue is urgent since there is no agreement on whether currently used criteria for medical interventions reflect the preferences of the general population, nor on whether these differ from patient preferences. In this study, respondents were asked to choose between two hypothetical scenarios of patients receiving a new treatment. The scenarios graphically represent treatment outcomes and patient characteristics. Their preferences were strongly and significantly affected by additional survival years, age at treatment, initial health condition, and the patient's lifestyle. The differences between patients and the general population appeared to be modest. However, apart from health gains, respondents thought the age of an individual as well as the cause and burden of disease were important factors in choosing which treatments should be provided to whom. This finding contradicts the assumptions underlying many of the procedures used to prioritize new medical interventions.

The aim of the **fifth** chapter was to explore the attendance to various information cues presented in the discrete-choice response tasks. A crucial assumption underpinning health measurement methods is that respondents pay equal attention to all information components presented in the response task. So far, no solid evidence has been found that respondents are fulfilling this condition. Eye-tracking was used to study the eye movements and fixations on specific information areas. This was done for seven discrete-choice response tasks comprising health-state descriptions. Videos of the eye movements of ten respondents were recorded and presented graphically. Frequencies were computed for length of fixation and number of fixations, and differences in attendance were demonstrated for particular health aspects in the tasks. All respondents completed the survey. Respondents were fixating on the left-sided health-state descriptions slightly longer than on the right-sided ones. Fatigue was not observed, as the time spent did not decrease in the final response tasks. The time spent on the tasks depended on the difficulty of the task and the amount of information presented. Eye-tracking proved to be a feasible method to study the process of paying attention and fixating on health-state descriptions in the discrete-choice response tasks.

SAMENVATTING

Het evalueren en meten van de subjectief ervaren gezondheidstoestand en de gezondheidsgerelateerde kwaliteit-van-leven is in toenemende mate relevant. Het meten hiervan omvat fysieke, mentale en sociale kenmerken. Er bestaan verschillende benaderingen voor het meten van zulke gezondheidstoestanden. In dit proefschrift gebruiken we waarderingsmethoden waarmee de algemene gezondheidstoestand van een persoon in een enkel maat of getal kan worden uitgedrukt. Zulke getallen worden waarden genoemd omdat ze de mate van ernst of kwaliteit van een gezondheidstoestand uitdrukken. Deze waarden zijn samengesteld uit gewichten die worden toegekend aan specifieke gezondheidsaspecten.

De waarderingsbenadering om te meten wordt in een aantal instrumenten gebruikt (bijv. EQ-5D, HUI-3, SF-6D, AQol). Voor het voorliggende proefschrift maakten we gebruik van het EQ-5D instrument dat uit vijf domeinen bestaat: mobiliteit, zelfzorg, dagelijkse activiteiten, pijn/klachten, en angst/depressie. Enkele methodologische nadelen van dit instrument waren de reden om van drie antwoordcategorieën per domein over te gaan naar vijf categorieën (EQ-5D-5L). Er is geen eenduidig bewijs gevonden dat het nieuwe formaat betere resultaten oplevert dan oorspronkelijke versie met drie categorieën (EQ-5D-3L).

Het doel van het **eerste hoofdstuk** was om een rechtstreekse vergelijking te maken van de EQ-5D-3L en de EQ-5D-5L bij het gebruik van een discreet keuze model. Het discreet keuze model is gebaseerd op moderne meettheorie en is vergelijkbaar met een frequent gebruikt nutsmodel uit de economische theorie. Bij toepassing van de discrete keuzetechniek worden de deelnemers gevraagd om een keuze te maken tussen twee of meer voorgelegde scenario's die worden beschreven in termen van specifieke domeinen (ook wel attributen genoemd) die uit verschillende niveaus (categorieën) bestaan. Er wordt daarbij vooral gelet op de consistentie en de logische rangschikking van de coëfficiënten die de niveaus van de attributen aangeven en op de verschillen in de waarden van de gezondheidstoestanden die verkregen worden op basis van deze coëfficiënten.

In het surveyonderzoek werden de antwoorden van een onderzoekspopulatie van 3,698 respondenten geanalyseerd en werden de uitkomsten van het toepassen van de EQ-5D-3L en EQ-5D-5L vergeleken met betrekking tot verschillende gezondheidstoestanden. In geen van beide versies was sprake van inconsistenties of van niet-logische rangschikking van niveaucoefficienten. Waarden afkomstig uit een grote representatieve steekproef voor de gehele Nederlandse bevolking waarin dezelfde manier van meten werd gebruikt met beide versies leverde wel kleine verschillen op in het gebruik van EQ-5D-3L en EQ-5D-5L. Naar onze mening kunnen dergelijke kleine verschillen het gevolg zijn van het bestaan van zelfs kleine verschillen in de formulering (taal) van de omschrijvingen of door het toepassen van een andere meetmethode.

Een karakteristiek kenmerk van waarderingsinstrumenten (bijv. SF-6D, EQ-5D, AQol) is dat ze in hun meetprocedures gebruik maken van statistische functies die alleen betrekking hebben op de hoofdeffecten van de afzonderlijke gezondheidsdomeinen. Dergelijke functies zijn gebaseerd op de vereenvoudigde aanname dat het totale effect van alle gezondheidsdomeinen tezamen gelijk is aan de som van de effecten van de afzonderlijke domeinen.

Maar gezondheidsdomeinen zijn vaak onderling verbonden en van elkaar afhankelijk. Eerdere studies (Feeny et al., 2002; Rowen et al., 2009) wezen uit dat de effecten van verschillende domeinen van gezondheidstoestanden niet alleen additief zijn maar dat er belangrijke interacties kunnen bestaan. Met andere woorden, dat de verslechtering van een ervaren gezondheidstoestand versterkt kan worden als twee verschillende gezondheidsproblemen op elkaar inwerken. Bij de ontwikkeling van het EQ-5D instrument werd in verschillende eerdere functies de N3 term gebruikt die de aanwezigheid van een ernstige score op een domein (sterkte van 3 in de EQ-5D-3L) aangeeft in tenminste een van de domeinen van de gezondheidstoestand. Echter, er werden geen aanwijzingen gevonden dat het opnemen van een dergelijke algemene interactie-term tot verbetering van de modeluitkomsten leidde.

Daarom is het onderzoek dat besproken wordt in het **tweede hoofdstuk** gericht op de vraag of het opnemen van interacties voor de EQ-5D-3L de uitkomsten van het model zou verbeteren en andere waarden voor de gezondheidstoestanden zou opleveren dan een functie die alleen uit hoofdeffecten bestaat. Het gebruik van een efficiënt onderzoekontwerp leidde tot het samenstellen van 400 paren van EQ-5D-3L gezondheidstoestanden. De vraag aan 4,000 personen uit de algemene populatie was welke van de twee gezondheidstoestanden beter was. De responses op deze paarsgewijze (discrete) keuze-taak werden geanalyseerd met gebruik van een conditional logit model waarbij ook de goodness-of-fit werd getest.

Globaal leidde de waardenfunctie voor de EQ-5D-3L waarin interacties waren opgenomen stelselmatig tot lagere uitkomsten dan het model met alleen hoofdeffecten. Hieruit kan afgeleid worden dat het gecombineerde effect van het bestaan van twee of meer gezondheidsproblemen tot een lagere waardering leidt dan de som van de afzonderlijke gezondheidsproblemen.

Over het algemeen worden in economische evaluaties de waarden van gezondheidstoestanden afgeleid in een representatieve steekproef afkomstig uit een cultureel en sociaal relevant geografisch gebied die vooral uit gezonde en uit tamelijk gezonde personen bestaat. Die keuze berust op het argument dat zij als belastingbetalers een afspiegeling zijn van de bevolking als totaal. Meer recent hebben onderzoekers in veel landen ervoor gekozen voor hun onderzoek gebruik te maken van waarderingen die uitgesproken worden door patienten. Deze keus berust op de redenering dat patienten beter geïnformeerd zijn over de consequenties van verslechterde gezondheid of beter

in staat zijn zich een voorstelling te maken van hypothetische gezondheidstoestanden, en daarom beter kunnen oordelen.

Het **derde hoofdstuk** laat verschillen zien tussen contrasterende steekproeven (mensen met ziekte-ervaring tegenover respondenten die momenteel gezond zijn) wat betreft de waarde die zij toekennen aan EQ-5D-5L gezondheidstoestanden. Hiervoor werd de response van 3,068 respondenten geanalyseerd. Er bleken verschillen te bestaan in de waardering tussen individuen met en zonder ziekte-ervaring, wat een verschil opleverde in de waardering van de gezondheidstoestanden door gezonde mensen en patienten. Daarmee ondersteunen de resultaten dat het belangrijk is om waardeoordelen te verzamelen onder mensen met ervaring van bepaalde gezondheidstoestanden of ziekten.

Hoofdstuk vier richt zich op de verschillen tussen twee steekproeven (gewone bevolking en patiënten), maar richt zich niet op de toepassing van het EQ-5D instrument. Het doel van dit onderzoek was om het belang te achterhalen van bepaalde criteria die gebruikt worden bij het toelaten van nieuwe medische behandelingen. Daarnaast richtte die onderzoek zich op het nagaan of er hierbij verschillen bestaan tussen de algemene bevolking en patiënten. Dit is van groot belang omdat er tot op heden geen overeenstemming bestaat over de vraag of de huidige criteria die worden gehanteerd voor het opnemen van medische ingrepen in de verzekerde zorg conform de voorkeuren van de gewone bevolking zijn, en of die verschillen van de voorkeuren van patienten. Voor dit onderzoek werden respondenten gevraagd een keuze te maken tussen twee hypothetische scenarios betreffende het aanbieden van een nieuwe behandeling aan patiënten. De scenarios behelsden een concreet beeld van de resultaten van de behandeling en kenmerken van de patienten.

De uitkomsten toonden aan dat de voorkeuren sterk beïnvloed worden door het aantal extra levensjaren, de leeftijd ten tijde van de behandeling, de staat van gezondheid in de uitgangsituatie, en de leefwijze van de patient. Daarbij lijken de verschillen tussen patienten en de gewone bevolking beperkt te zijn. Maar afgezien van gezondheidsverbetering meenden de respondenten dat de leeftijd van de betreffende persoon, alsmede de oorzaak en de mate van het lijden aan de ziekte, van belang moeten zijn bij de beslissing welke behandeling aan wie aangeboden zou moeten worden. Onze resultaten verschillen derhalve van de aannamen die ten grondslag liggen aan veel van de procedures die momenteel in gebruik zijn bij de prioritering van nieuwe medische ingrepen.

Het **vijfde hoofdstuk** had tot doel om te bepalen hoe respondenten reageren op verschillende informatie-elementen in de aan hen voorgelegde taken in het discrete-keuze taken. Een wezenlijke aanname die ten grondslag ligt aan alle methoden die gezondheidstoestanden waarderen is dat de respondenten aan alle informatieonderdelen in de voorgelegde waarderingstaken evenveel aandacht geven. Tot dusver bestaat er echter geen solide bewijs dat deze aanname ondersteunt. Het volgen van oogbewegingen (eye-tracking) en het richten van de blik op specifieke

delen van de informatie werd gebruikt om dit te onderzoeken. Deze methode werd toegepast op zeven discrete keuze response taken met betrekking tot beschrijvingen van gezondheidstoestanden. Video-opnamen werden gemaakt van de oogbewegingen van tien respondenten en grafisch weergegeven. Frequenties werden berekend voor de tijdsduur van een blik op informatieonderdelen en van het aantal keren dat de blik werd vastgehouden. Dat leidde tot het identificeren van verschillen in aandacht voor bepaalde instructies en gezondheidsaspecten in de taken. Alle respondenten maakten het experiment af. Ze bleken iets langer aandacht te geven aan beschrijvingen van gezondheidstoestanden die links stonden dan aan degene die aan de rechterkant van het beeldscherm waren beschreven. Uit het feit dat de tijdsduur tijdens de laatste discrete keuze niet korter was dan bij de eerdere, wordt afgeleid dat er geen aanwijzing was voor het optreden van vermoeidheid bij de respondenten. De hoeveelheid tijd die aan de taken werd gewijd was afhankelijk van de moeilijkheidsgraad van de opdracht en de hoeveelheid gepresenteerde informatie. Hieruit blijkt dat het volgen van oogbewegingen een goede aanpak is om te onderzoeken hoe het proces van aandacht geven aan en het concentreren op beschrijvingen van gezondheidstoestanden als onderdeel van discrete keuze response taken verloopt.

ACKNOWLEDGEMENTS

Today I am starting to write the acknowledgements sections, and the thought that immediately appears in my head: “I could not even imagine how I would feel when this day happens”. How do all PhD candidates feel when they are at home stretch, while writing the Acknowledgements? Probably, the memories of good and bad moments, exciting or stressful events are popping up. What would you feel: relief, excitement, happiness, worries?

I am writing this section having already moved from the Netherlands to Switzerland. Now sitting towards the lake, I am reflecting on the past 3 years of PhD life in the Netherlands. It is not hard to realize that there were definitely plenty of moments to remember, full of excitement and joy, as well as despair and hard-working. Now, in retrospective, I understand that such moments are worth experiencing. Without downs it is not possible to appreciate ups (then again downs and even higher ups), and being thankful for that.

First of all, I would like to express my thankfulness to my supervisor, Dr. Paul Krabbe. Apart from being my supervisor and leader in science and research, he also became my guide in life (and wine, and music). Thanks for all your support and trust in me. Thanks for giving clear instructions and willingness to help. Thanks for all interesting talks and stories to discuss. Especially, I will always remember the pleasant evening in your house in Zeist with your wife Anna, and colleagues Ruslan and Ahmad. There is such welcoming yet calming atmosphere in your house, surrounded by nature. Not to forget the beautiful music we were listening, and the elegant meal prepared by Anna and you. Special thanks for being flexible always (so I could travel sometimes) and especially at the last stage of my PhD when I was allowed to finish my PhD project abroad.

Additionally, I would like to express my gratitude to Professor Erik Buskens for all great help with writing the manuscripts and organization of the last steps before the thesis was complete. I remember that crucial telephone call, when you were able to discuss with me the final steps of the PhD while biking. Thanks for the possibility to find time for me.

I would like to thank Karin Vermeulen for giving useful comments and all help during writing the difficult papers. Additionally, thanks for the great time in Philadelphia after the ISOQOL conference, when we walked in the city and visited the museum of Rodin.

Thanks a lot to Aukje van der Zee and Roelian Geuze for always being helpful to me in all arrangements for meetings, conferences, and foreign trips. Additional and substantial thanks for helping me in my eye-tracking study. Apart from work, thank you so much for very lively discussions on traveling, so I could find inspiration from your stories, impressions, and happy smiles.

I would like to thank Marijke Hanania for all the support and help provided with organization of the foreign trips and department events. I would also like to thank Thea van Asselt for being great help with writing the manuscript and always being able to

reply on time. Additional thanks for being the participant in the trial of my eye-tracking device for the study.

I express my gratitude to all the members of EuroQol group for funding this PhD project and letting me conduct it. Additionally, thanks for organizing high-end and very informative events in beautiful places. The annual plenary meeting in Krakow was the first scientific event of high volume and standard I ever attended. I remember how nervous I was to present my poster, but thankfully it ended well. However, the excitement to present at EuroQol meetings did not leave me from one year to the other. I am expressing genuine gratitude for all useful and critical remarks that have been made for my work. It helped me to improve the papers so much, that both manuscripts presented at the meetings were eventually accepted in high ranked journals for publication.

Of course, apart from the scientific part, there was a great social environment, where I met many interesting people. I would like to thank Fred, Joey, Jessica, Milad, and Gerben for such great time spent together on the conferences and supporting me in preparation of my presentations. Moreover, we had a lot of funny stories told, a lot of sightseeing made, and we laughed a lot, which made my conference trips even more memorable.

Moving the memories back to the office in Groningen, the agenda was not only full of writing and statistics. I met a lot of great people there, whom I should thank for making life brighter. Alicja, Natalia, Sara - thanks for all beautiful trips we made and nice time we spent together. Additional thanks for being such great companions on the music festivals and gym trainings we had. Ruslan, Grigory, Elnaz and Eliza - thank you so much for being a great lunch team, when we could talk about everything and make a small break in the middle of the day to refresh our minds. Jacobien, thanks a lot for being such a great officemate, although for such a short time. Ahmad, thanks a lot for always being so kind and in good mood, and for all the fruits you were giving to your starving colleagues. Reinder, I would like to express many thanks for all those supporting and inspiring conversations at the end of the day. When I was sitting alone in the office, exhausted from work, you appeared and made me laugh. Petra, Joyce, Tugs and Ariuntaya, thanks for being so friendly and creating such good atmosphere in the office. Mohammedreza and Kebede, thanks for being such great and understanding office mates. Neda, your cookies were super delicious!

These 3 years of my PhD track were not full of science only, but were also full of great free time events. Bootcamp trainings, wall climbing, wine and cheese tasting, potluck dinners, Halloween parties, camp trips, alternative city tours, pub crawling, gala dinners, lectures… This is just a short list of all activities organized by the PhD organization Gopher. Thanks to all who organized all these great activities, and those who participated. These events were truly unforgettable. Special thanks to my greatest friends I managed to find in Groningen – Zhenya, Julius, Oleg, Artur – you were true support for me these years. It is not only the fact that we all could speak the same language (Russian). It is all about understanding and unity, creative spirit and willingness to join me in my crazy

ideas, which made time spent with you truly amazing. Playing pool and board games, parties, barbeques, day and weekend trips, sport events and just cosy evenings together made me feel like at home. A place where I can just be myself.

Work and social life are important parts of the life, but what is truly important is family. I am thankful to my parents, grandparents and sister for supporting me. Although they were not physically with me, I always felt that they are not thousands kilometres away, but right here. I could always count on them and find encouragement. Thanks for always being there for me.

Special words I would like to say to my husband, Christophe, for all support and care. We met each other in the Netherlands during the period of my PhD and became a family not long ago. This event was another gift, the life in the Netherlands has brought me. In very hard times he is always here, ready to listen and motivate. In addition, thanks for all the strength and patience, although sometimes it is not easy to handle all my worries.

Walther, Bettina, Mathias, Rachel, Ulrich, and Maxim, thank you very much for becoming my second family. The warmth and understanding I got from you is incredible. Having you all around, helped me to finally feel home after a lot of traveling from one country to the other (Russia to the Netherlands and back, Netherlands to Switzerland and back) and to be a part of a very strong, very connected family.

And last but not least, I would like to express my thankfulness to the Netherlands, and the city of Groningen in particular. This is not a joke. This small but cosy city accepted me well and opened the doors for my future personal and career development. In spite of the weather, which is often the topic of conversations and complaints, Groningen has its charm and warmth. It should be mentioned that the amount of cafes, bars and events is overwhelming and helped to brighten up the free time. Moreover, not to forget: in what other country would I be able to live in a houseboat? A house on water! Is that not a great experience? This experience adds up to all the other experiences I had in the Netherlands these years, forming up one big adventure. Life is a great adventure, FULL of ups and downs, life-changing events and usual routines. However, the most important is that life has to be FULL. So, I was happy to have my adventure in the Netherlands, and thanks to you all for joining it.

Мама и папа, бабушка и дедушка, Катюшка!
Спасибо вам всем за поддержку и терпение. Спасибо и за то, что вы меня очень многому научили: бабушка-трудолюбию и ответственности; дедушка-изобретательности и умению преодолевать препятствия; папа-быть сильной и упорной; мама-ничего не бояться и идти с высоко поднятой головой. Катя научила меня, что главное-семья и любовь есть. Так что вы все, каждый по-своему, помогли мне дойти до конца этого пути. Получение степени и эта книга - заслуга нас всех, и спасибо вам за это!

CURRICULUM VITAE

Anna Nicolet (Selivanova) was born on 27 March 1992 in the city of Ryazan in Russian Federation. This city is located not far from Moscow, which is 200 kilometers away. In 1999, Anna started the study in Gymnasium 5, which specializes in foreign languages and literature. In Gymnasium 5 Anna learnt English, German and Latin. At the end of the study, in 2009, Anna received the silver medal for excellent achievements and successfully passed the state exams to be enrolled in the university. Pursuing the need for challenges, Anna chose to move to Moscow to get a Bachelor's degree in the National Research University – Higher School of Economics, which is considered one of the best universities in Russia with major in Economic science. During her study in the university, Anna also took a job in the fitness industry as a fitness instructor (fitness clubs "The Star" and "Fitness Empire") and additionally in the bank business as a consultant (Metallinvestbank). As part of the Bachelor's degree course, completing an internship was obligatory. Therefore, Anna was accepted as an intern in the Ministry of Economic Development of the Russian Federation in the department of Physical Culture and Sport. During the internship and fitness-related job, she developed the interest in sport and healthy lifestyle. The Bachelor's degree was finished by 2013, whereby the need for a new challenge appeared. Anna moved to the Netherlands for the Master's degree in Health Economics. Health Economics specialization in Erasmus School of Economics (Rotterdam) was chosen based on the interest in healthy behavior. Anna finished her Master's degree in Rotterdam within 1 year, successfully defended, and then published her Master's thesis on the relationship between healthy behaviors and health outcomes among older adults in Russia. After finishing the Master's degree, Anna returned to Russia and after half a year, she was accepted on the PhD position in the University of Groningen, University Medical Center Groningen in the Netherlands. Anna started the 3-year PhD track in 2015 in Epidemiology, Health Technology Assessment unit. Her PhD thesis explored the health-state valuation using discrete choice modelling, which was completed by September 2018.

In February 2018 Anna married and moved to Switzerland, canton Fribourg. Currently, she is working as Associate Scientist in Philip Morris International, Neuchâtel, Switzerland. She is still passionate about travelling and at the age of 26 visited over 26 countries.

SHARE Previous Dissertations

Research Institute SHARE

This thesis is published within the **Research Institute SHARE** (Science in Healthy Ageing and healthcaRE) of the University Medical Center Groningen / University of Groningen. Further information regarding the institute and its research can be obtained from our internet site: http://www.share.umcg.nl/

More recent theses can be found in the list below.
((co-) supervisors are between brackets)

2018

Arifin B
Distress and health-related quality of life in Indonesian Type 2 diabetes mellitus outpatients
(prof MJ Postma, dr PJM Krabbe, dr J Atthobari)

Zakiyah N
Women's health from a global economic perspective
(prof MJ Postma, dr ADI van Asselt)

Metting, EI
Development of patient centered management of asthma and COPD in primary care
(prof T van der Molen, prof R Sanderman, dr JWH Kocks)

Suhoyo Y
Feedback during clerkships: the role of culture
(prof JBM Kuks, prof J Cohen-Schotanus, dr J Schönrock-Adema)

Veen HC van der
Articulation issues in total hip arthroplasty
(prof SK Bulstra, dr JJAM van Raay, dr IHF Reininga, dr I van den Akker-Scheek)

Elsenburg LK
Adverse life events and overweight in childhood, adolescence and young adulthood
(prof AC Liefbroer, dr N Smidt)

't Hoen EFM
Practical applications of the flexibilities of the agreement on trade-related aspects of intellectual property rights; lessons beyond HIV for access to new essential medicines
(prof HV Hogerzeil, prof BCA Toebes)

Stojanovska V
Fetal programming in pregnancy-associated disorders; studies in novel preclinical models
(prof SA Scherjon, dr T Plösch)

Eersel MEA van
The association of cognitive performance with vascular risk factors across adult life span
(prof JPJ Slaets, dr GJ Izaks, dr JMH Joosten)

Rolfes L
Patient participation in pharmacovigilance
(prof EP van Puijenbroek, prof K Taxis, dr FPAM van Hunsel)

Brandenbarg D
The role of the general practitioner in the care for patients with colorectal cancer
(prof MY Berger, prof GH de Bock, dr AJ Berendsen)

Oldenkamp M
Caregiving experiences of informal caregivers; the importance of characteristics of the informal caregiver, care recipient, and care situation
(prof RP Stolk, prof M Hagedoorn, prof RPM Wittek, dr N Smidt)

Kammen K van
Neuromuscular control of Lokomat guided gait; evaluation of training parameters
(prof LHV van der Woude, dr A den Otter, dr AM Boonstra, dr HA Reinders-Messelink)

Hornman J
Stability of development and behavior of preterm children
(prof SA Reijneveld, prof AF Bos, dr A de Winter)

Vries, YA de
Evidence-b(i)ased psychiatry
(prof P de Jonge, dr AM Roest)

Smits KPJ
Quality of prescribing in chronic kidney disease and type 2 diabetes
(prof P denig, prof GJ Navis, prof HJG Bilo, dr GA Sidorenkov)

Zhan Z
Evaluation and analysis of stepped wedge designs; application to colorectal cancer follow-up
(prof GH de Bock, prof ER van den Heuvel)

Hoeve Y ten
From student nurse to nurse professional
(prof PF Roodbol, prof S Castelein, dr GJ Jansen, dr ES Kunnen)

Ciere Y
Living with chronic headache
(prof R Sanderman, dr A Visser, dr J Fleer)

Borkulo CD van
Symptom network models in depression research; from methodological exploration to clinical application
(prof R Schoevers, prof D Borsboom, dr L Boschloo, dr LJ Waldorp)

For more 2018 and earlier theses visit our website